entropy

MDPI

Article

Finite-Length Analyses for Source and Channel Coding on Markov Chains†

Masahito Hayashi [1,2,3,4,*,‡] **and Shun Watanabe** [5,‡]

1 Shenzhen Institute for Quantum Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
2 Graduate School of Mathematics, Nagoya University, Nagoya 464-8602, Japan
3 Center for Quantum Computing, Peng Cheng Laboratory, Shenzhen 518000, China
4 Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore 117542, Singapore
5 Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, Koganei-shi, Tokyo 184-8588, Japan; shunwata@cc.tuat.ac.jp
* Correspondence: hayashi@sustech.edu.cn or masahito@math.nagoya-u.ac.jp
† This paper is an extended version of the conference paper presented at the 51st Allerton Conference and 2014 Information Theory and Applications Workshop, San Diego, CA, USA, 9–14 February 2014.
‡ These authors contributed equally to this work.

Received: 9 March 2020; Accepted: 4 April 2020; Published: 18 April 2020

Abstract: We derive finite-length bounds for two problems with Markov chains: source coding with side-information where the source and side-information are a joint Markov chain and channel coding for channels with Markovian conditional additive noise. For this purpose, we point out two important aspects of finite-length analysis that must be argued when finite-length bounds are proposed. The first is the asymptotic tightness, and the other is the efficient computability of the bound. Then, we derive finite-length upper and lower bounds for the coding length in both settings such that their computational complexity is low. We argue the first of the above-mentioned aspects by deriving the large deviation bounds, the moderate deviation bounds, and second-order bounds for these two topics and show that these finite-length bounds achieve the asymptotic optimality in these senses. Several kinds of information measures for transition matrices are introduced for the purpose of this discussion.

Keywords: channel coding; Markov chain; finite-length analysis; source coding

1. Introduction

In recent years, finite-length analyses for coding problems have been attracting considerable attention [1]. This paper focuses on finite-length analyses for two representative coding problems: One is source coding with side-information for Markov sources, i.e., the Markov–Slepian–Wolf problem on the system X^n with full side-information Y^n at the decoder, where only the decoder observes the side-information and the source and the side-information are a joint Markov chain. The other is channel coding for channels with Markovian conditional additive noise. Although the main purpose of this paper is finite-length analyses, we also present a unified approach we developed to investigate these topics including asymptotic analyses. Since this discussion is spread across a number of subtopics, we explain them separately in the Introduction.

1.1. Two Aspects of Finite-Length Analysis

We explain the motivations of this research by starting with two aspects of finite-length analysis that must be argued when finite-length bounds are proposed. For concreteness, we consider channel

coding here even though the problems treated in this paper are not restricted to channel coding. To date, many types of finite-length achievability bounds have been proposed. For example, Verdú and Han derived a finite-length bound by using the information-spectrum approach in order to derive the general formula [2] (see also [3]), which we term as the information-spectrum bound. One of the authors and Nagaoka derived a bound (for the classical-quantum channel) by relating the error probability to binary hypothesis testing [4] (Remark 15) (see also [5]), which we refer to as the hypothesis-testing bound. Polyanskiy et. al. derived the random coding union (RCU) bound and the dependence testing (DT) bound [1] (a bound slightly looser (coefficients are worse) than the DT bound can be derived from the hypothesis-testing bound of [4]). Moreover, Gallager's bound [6] is known as an efficient bound to derive the exponentially decreasing rate.

Here, we focus on two important aspects of finite-length analysis:

(A1) Computational complexity for the bound and
(A2) Asymptotic tightness for the bound.

Both aspects are required for the bound in finite-length analysis as follows. As the first aspect, we consider the computational complexity for the bound. For the BSC (binary symmetric channel), the computational complexity of the RCU bound is $O(n^2)$, and that of the DT bound is $O(n)$ [7]. However, the computational complexities of these bounds are much larger for general DMCs (discrete memoryless channels) or channels with memory. It is known that the hypothesis testing bound can be described as a linear programming problem (e.g., see [8,9] (in the the case of a quantum channel, the bound is described as a semi-definite programming problem)) and can be efficiently computed under certain symmetry. However, the number of variables in the linear programming problem grows exponentially with the block length, and it is difficult to compute in general. The computation of the information-spectrum bound depends on the evaluation of the tail probability. The hypothesis testing bound gives a tighter bound than the information-spectrum bound, as pointed out by [8], and the computational complexity of the former is much smaller than that of the latter. However, the computation of the tail probability continues to remain challenging unless the channel is a DMC. For DMCs, the computational complexity of Gallager's bound is $O(1)$ since the Gallager function is an additive quantity for DMCs. However, this is not the case if there is a memory (the Gallager bound for finite-state channels was considered in [10] (Section 5.9), but a closed form expression for the exponent was not derived). Consequently, no efficiently computable bound currently exists for channel coding with Markov additive noise. The situation is the same for source coding with side-information.

Since the actual computation time may depend on the computational resource we can use for numerical experiment, it is not possible to provide a concrete requirement of computational complexity. However, in order to conduct a numerical experiment for a meaningful blocklength, it is reasonable to require the computational complexity to be, at most, a polynomial order of the blocklength n.

Next, let us consider the second aspect, i.e., asymptotic tightness. Thus far, three kinds of asymptotic regimes have been studied in information theory [1,11–16]:

- A large deviation regime in which the error probability ε asymptotically behaves as e^{-nr} for some $r > 0$;
- A moderate deviation regime in which ε asymptotically behaves as $e^{-n^{1-2t}r}$ for some $r > 0$ and $t \in (0, 1/2)$; and
- A second-order regime in which ε is a constant.

We shall claim that a good finite-length bound should be asymptotically optimal for at least one of the above-mentioned three regimes. In fact, the information-spectrum bound, the hypothesis-testing bound, and the DT bound are asymptotically optimal in both the moderate deviation and second-order regimes, whereas the Gallager bound is asymptotically optimal in the large deviation regime and the RCU bound asymptotically optimal in all the regimes (Both the Gallager and RCU bounds are asymptotically optimal in the large deviation regime only up to the critical rate). Recently, for

DMCs, Yang and Meng derived an efficiently computable bound for low-density parity check (LDPC) codes [17], which is asymptotically optimal in both the moderate deviation and second-order regimes.

1.2. Main Contribution for Finite-Length Analysis

We derive the finite-length achievability bounds on the problems by basically using the exponential-type bounds (for channel coding, it corresponds to the Gallager bound.). In source coding with side-information, the exponential-type upper bounds on error probability $\bar{\mathrm{P}}_{\mathrm{e}}(M_n)$ for a given message size M_n are described by using the conditional Rényi entropies as follows (cf. Lemmas 14 and 15):

$$\bar{\mathrm{P}}_{\mathrm{e}}(M_n) \leq \inf_{-\frac{1}{2}\leq\theta\leq 0} M_n^{\frac{\theta}{1+\theta}} e^{-\frac{\theta}{1+\theta} H^{\uparrow}_{1+\theta}(X^n|Y^n)} \tag{1}$$

and:

$$\bar{\mathrm{P}}_{\mathrm{e}}(M_n) \leq \inf_{-1\leq\theta\leq 0} M_n^{\theta} e^{-\theta H^{\downarrow}_{1+\theta}(X^n|Y^n)}. \tag{2}$$

Here, X^n is the information to be compressed and Y^n is the side-information that can be accessed only by the decoder. $H^{\uparrow}_{1+\theta}(X^n|Y^n)$ is the conditional Rényi entropy introduced by Arimoto [18], which we shall refer to as the upper conditional Rényi entropy (cf. (12)). On the other hand, $H^{\downarrow}_{1+\theta}(X^n|Y^n)$ is the conditional Rényi entropy introduced in [19], which we shall refer to as the lower conditional Rényi entropy (cf. (7)). Although there are several other definitions of conditional Rényi entropies, we only use these two in this paper; see [20,21] for an extensive review on conditional Rényi entropies.

Although the above-mentioned conditional Rényi entropies are additive for i.i.d. random variables, they are not additive for joint Markov chains over X^n and Y^n, for which the derivation of finite-length bounds for Markov chains are challenging. Because it is generally not easy to evaluate the conditional Rényi entropies for Markov chains, we consider two assumptions in relation to transition matrices: the first assumption, which we refer to as non-hidden, is that the Y-marginal process is a Markov chain, which enables us to derive the single-letter expression of the conditional entropy rate and the lower conditional Rényi entropy rate; the second assumption, which we refer to as strongly non-hidden, enables us to derive the single-letter expression of the upper conditional Rényi entropy rate; see Assumptions 1 and 2 of Section 2 for more detail (Indeed, as explained later, our result on the data compression can be converted to a result on the channel coding for a specific class of channels. Under this conversion, we obtain certain assumptions for channels. As explained later, these assumptions for channels are more meaningful from a practical point of view.) . Under Assumption 1, we introduce the lower conditional Rényi entropy for transition matrices $H^{\downarrow,W}_{1+\theta}(X|Y)$ (cf. (47)). Then, we evaluate the lower conditional Rényi entropy for the Markov chain in terms of its transition matrix counterpart. More specifically, we derive an approximation:

$$H^{\downarrow}_{1+\theta}(X^n|Y^n) = nH^{\downarrow,W}_{1+\theta}(X|Y) + O(1), \tag{3}$$

where an explicit form of the $O(1)$ term is also derived. Using the evaluation (2) with this evaluation, we obtain finite-length bounds under Assumption 1. Under a more restrictive assumption, i.e., Assumption 2, we also introduce the upper conditional Rényi entropy for a transition matrix $H^{\uparrow,W}_{1+\theta}(X|Y)$ (cf. (55)). Then, we evaluate the upper Rényi entropy for the Markov chain in terms of its transition matrix counterpart. More specifically, we derive an approximation:

$$H^{\uparrow}_{1+\theta}(X^n|Y^n) = nH^{\uparrow,W}_{1+\theta}(X|Y) + O(1), \tag{4}$$

where an explicit form of the $O(1)$ term is also derived. Using the evaluation (1) with this evaluation, we obtain finite-length bounds that are tighter than those obtained under Assumption 1. It should

be noted that, without Assumption 1, even the conditional entropy rate is challenging to evaluate. For evaluation of the conditional entropy rate of the X process given the Y process, the assumption of the X process being Markov seems to be not helpful. This is the reason why we consider the Y process being Markov instead of the X process being Markov in this paper.

We also derive converse bounds by using the change of measure argument for Markov chains developed by the authors in the accompanying paper on information geometry [22,23]. For this purpose, we further introduce two-parameter conditional Rényi entropy and its transition matrix counterpart (cf. (18) and (59)). This novel information measure includes the lower conditional Rényi entropy and the upper conditional Rényi entropy as special cases. We clarify the relation among bounds based on these quantities by numerically calculating the upper and lower bounds for the optimal coding rate in source coding with a Markov source in Section 3.7. Owing to the second aspect (A2), this calculation shows that our finite-length bounds are very close to the optimal value. Although this numerical calculation contains a case with a very large size $n = 1 \times 10^5$, its calculation is not as difficult because the calculation complexity behaves as $O(1)$. That is, this calculation shows the advantage of the first aspect (A1).

Here, we would like to remark about the terminologies because there are a few ways to express exponential-type bounds. In statistics or large deviation theory, we usually use the cumulant generating function (CGF) to describe exponents. In information theory, we employ the Gallager function or the Rényi entropies. Although these three terminologies are essentially the same quantity and are related by the change of variables, the CGF and the Gallager function are convenient for some calculations because of their desirable properties such as convexity. On the other hand, the minimum entropy and collision entropy are often used as alternative information measures of Shannon entropy in the community of cryptography. Since the Rényi entropies are a generalization of the minimum entropy and collision entropy, we can regard the Rényi entropies as information measures. The information theoretic meaning of the CGF and the Gallager function are less clear. Thus, the Rényi entropies are intuitively familiar to the readers' of this journal. The Rényi entropies have an additional advantage in that two types of bounds (e.g., (152) and (161)) can be expressed in a unified manner. Therefore, we state our main results in terms of the Rényi entropies, whereas we use the CGF and the Gallager function in the proofs. For the readers' convenience, the relation between the Rényi entropies and corresponding CGFs are summarized in Appendices A and B.

1.3. Main Contribution for Channel Coding

An intimate relationship is known to exist between channel coding and source coding with side-information (e.g., [24–26]). In particular, for an additive channel, the error probability of channel coding by a linear code can be related to the corresponding source coding problem with side-information [24]. Chen et. al. also showed that the error probability of source coding with side-information by a linear encoder can be related to the error probability of a dual channel coding problem and vice versa [27] (see also [28]). Since these dual channels can be regarded as additive channels conditioned on state information, we refer to these channels as conditional additive channels (In [28], we termed these channels general additive channels, but we think "conditional" more suitably describes the situation.). In this paper, we mainly discuss a conditional additive channel, in which the additive noise is operated subject to a distribution conditioned on additional output information. Then, we convert our obtained results of source coding with side-information to the analysis on conditional additive channels. That is, using the aforementioned duality between channel coding and source coding with side-information enables us to evaluate the error probability of channel coding for additive channels. Then, we derive several finite-length analyses on additive channels.

For the same reason as source coding with side-information, we make two assumptions, Assumptions 1 and 2, on the noise process of a conditional additive channel. In this context, Assumption 1 means that the marginal system Y^n deciding the behavior of the additive noise X^n is a Markov chain. It should be noted that the Gilbert–Elliott channel [29,30] with state information

available at the receiver can be regarded as a conditional additive channel such that the noise process is a Markov chain satisfying both Assumptions 1 and 2 (see Example 6). Thus, we believe that Assumptions 1 and 2 are quite reasonable assumptions.

In fact, our analysis is applicable for a broader class of channels known as regular channels [31]. The class of regular channels includes conditional additive channels as a special case, and it is known as a class of channels that are similarly symmetrical. To show it, we propose a method to convert a regular channel into a conditional additive channel such that our treatment covers regular channels. Additionally, we show that the BPSK (binary phase shift keying)-AWGN (additive white Gaussian noise) channel is included in conditional additive channels.

1.4. Asymptotic Bounds and Asymptotic Tightness for Finite-Length Bounds

We present asymptotic analyses of the large and moderate deviation regimes by deriving the characterizations (for the large deviation regime, we only derive the characterizations up to the critical rate) with the use of our finite-length achievability and converse bounds, which implies that our finite-length bounds are tight in both of these deviation regimes. We also derive the second-order rate. Although this rate can be derived by the application of the central limit theorem to the information-spectrum bound, the variance involves the limit with respect to the block length because of memory. In this paper, we derive a single-letter form of the variance by using the conditional Rényi entropy for transition matrices (An alternative way to derive a single-letter characterization of the variance for the Markov chain was shown in [32] (Lemma 20). It should also be noted that a single-letter characterization can be derived by using the fundamental matrix [33]. The single-letter characterization of the variance in [12] (Section VII) and [11] (Section III) contains an error, which is corrected in this paper.).

As we will see in Theorems 11–14 and 22–25, our asymptotic results have the same forms as the counterparts of the i.i.d. case (cf. [1,6,11–14]) when the information measures for distributions in the i.i.d. case are replaced by the information measures for the transition matrices introduced in this paper.

We determine the asymptotic tightness for finite-length bounds by summarizing the relation between the asymptotic results and the finite-length bounds in Table 1. The table also describes the computational complexity of the finite-length bounds. "Solved*" indicates that those problems are solved up to the critical rates. "Ass. 1" and "Ass. 2" indicate that those problems are solved either under Assumption 1 or Assumption 2. "$O(1)$" indicates that both the achievability and converse parts of those asymptotic results are derived from our finite-length achievability bounds and converse bounds whose computational complexities are $O(1)$. "Tail" indicates that both the achievability and converse parts of those asymptotic results are derived from the information-spectrum-type achievability bounds and converse bounds of which the computational complexities depend on the computational complexities of the tail probabilities.

In general, the exact computations of tail probabilities are difficult, although they may be feasible for a simple case such as an i.i.d. case. One way to compute tail probabilities approximately is to use the Berry–Esséen theorem [34] (Theorem 16.5.1) or its variant [35]. This direction of research is still ongoing [36,37], and an evaluation of the constant was conducted [37], although its tightness has not been clarified. If we can derive a tight Berry–Esséen-type bound for the Markov chain, this would enable us to derive a finite-length bound that is asymptotically tight in the second-order regime. However, the approximation errors of Berry–Esséen-type bounds converge only in the order of $1/\sqrt{n}$ and cannot be applied when ε is rather small. Even in cases in which the exact computations of tail probabilities are possible, the information-spectrum-type bounds are looser than the exponential type bounds when ε is rather small, and we need to use appropriate bounds depending on the size of ε. In fact, this observation was explicitly clarified in [38] for random number generation with side-information. Consequently, we believe that our exponential-type finite-length bounds are very useful. It should be also noted that, for source coding with side-information and channel coding for regular channels, even the first-order results have not been revealed as far as the authors know, and

they are clarified in this paper (General formulae for those problems were known [2,3], but single-letter expressions for Markov sources or channels were not clarified in the literature. For the source coding without side-information, the single-letter expression for entropy rate of Markov source is well known (e.g., see [39]).).

Table 1. Summary of asymptotic results and finite-length bounds to derive asymptotic results under Assumptions 1 and 2, which are abbreviated to Ass. 1 and Ass. 2.

Problem	First-Order	Large Deviation	Moderate Deviation	Second-Order
SC with SI	Solved (Ass. 1)	Solved* (Ass. 2) $O(1)$	Solved (Ass. 1), $O(1)$	Solved (Ass. 1) Tail
CC for Conditional Additive Channels	Solved (Ass. 1)	Solved* (Ass. 2) $O(1)$	Solved (Ass. 1) $O(1)$	Solved (Ass. 1) Tail

1.5. Related Work on Markov Chains

Since related work concerning the finite-length analysis is reviewed in Section 1.1, we only review work related to the asymptotic analysis here. Some studies on Markov chains for the large deviation regime have been reported [40–42]. The derivation in [40] used the Markov-type method. A drawback of this method is that it involves a term that stems from the number of types, which does not affect the asymptotic analysis, but does hurt the finite-length analysis. Our achievability is derived by following a similar approach as in [41,42], i.e., the Perron–Frobenius theorem, but our derivation separates the single-shot part and the evaluation of the Rényi entropy, and thus is more transparent. Furthermore, the converse part of [41,42] is based on the Shannon–McMillan–Breiman limiting theorem and does not yield finite-length bounds.

For the second-order regime, Polyanskiy et. al. studied the second-order rate (dispersion) of the Gilbert–Elliott channel [43]. Tomamichel and Tan studied the second-order rate of channel coding with state information such that the state information may be a general source and derived a formula for the Markov chain as a special case [32]. Kontoyiannis studied the second-order variable length source coding for the Markov chain [44]. In [45], Kontoyiannis and Verdú derived the second-order rate of lossless source coding under the overflow probability criterion.

For channel coding of the i.i.d. case, Scarlett et al. derived a saddle-point approximation, which unifies all three regimes [46,47].

1.6. Organization of the Paper

In Section 2, we introduce the information measures and their properties that will be used in Sections 3 and 4. Then, source coding with side-information and channel coding is discussed in Sections 3 and 4, respectively. As we mentioned above, we state our main result in terms of the Rényi entropies, and we use the CGFs and the Gallager function in the proofs. We explain how to cover the continuous case in Remarks 1 and 5. In Appendices A and B, the relation between the Rényi entropies and corresponding CGFs are summarized. The relation between the Rényi entropies and the Gallager function are explained as necessary. Proofs of some technical results are also provided in the remaining Appendices.

1.7. Notations

For a set $\mathcal{X}$, the set of all distributions on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$. The set of all sub-normalized non-negative functions on $\mathcal{X}$ is denoted by $\bar{\mathcal{P}}(\mathcal{X})$. The cumulative distribution function of the standard Gaussian random variable is denoted by:

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right] dx. \tag{5}$$

Throughout the paper, the base of the logarithm is the natural base e.

2. Information Measures

Since this paper discusses the second-order tightness, we need to discuss the central limit theorem for the Markov process. For this purpose, we usually employ advanced mathematical methods from probability theory. For example, the paper [48] (Theorem 4) showed the Markov version of the central limit theorem by using a martingale stopping technique. Lalley [49] employed the regular perturbation theory of operators on the infinite-dimensional space [50] (Chapter 7, #1, Chapter 4, #3, and Chapter 3, #5). The papers [51,52] and [53] (Lemma 1.5 of Chapter 1) employed the spectral measure, while it is hard to calculate the spectral measure in general even in the finite-state case. Further, the papers [36,51,54,55] showed the central limit theorem by using the asymptotic variance, but they did not give any computable expression of the asymptotic variance without the infinite sum. In summary, to derive the central limit theorem with the variance of a computable form, these papers needed to use very advanced mathematics beyond calculus and linear algebra.

To overcome the difficulty of the Markov version of the central limit theorem, we employed the method used in our recent paper [23]. The paper [23] employed the method based on the cumulant generating function for transition matrices, which is defined by the Perron eigenvalue of a specific non-negative-entry matrix. Since a Perron eigenvalue can be explained in the framework of linear algebra, the method can be described with elementary mathematics. To employ this method, we need to define the information measure in a way similar to the cumulant generating function for transition matrices. That is, we define the information measures for transition matrices, e.g., the conditional Rényi entropy for transition matrices, etc, by using Perron eigenvalues.

Fortunately, these information measures for transition matrices are very useful even for large deviation-type evaluation and finite-length bounds. For example, our recent paper [23] derived finite-length bounds for simple hypothesis testing for the Markov chain by using the cumulant generating function for transition matrices. Therefore, using these information measures for transition matrices, this paper derives finite-length bounds for source coding and channel coding with Markov chains and discusses their asymptotic bounds with large deviation, moderate deviation, and the second-order type.

Since they are natural extensions of information measures for single-shot setting, we first review information measures for the single-shot setting in Section 2.1. Next, we introduce information measures for transition matrices in Section 2.2. Then, we show that information measures for Markov chains can be approximated by information measures for transition matrices generating those Markov chains in Section 2.3.

2.1. Information Measures for the Single-Shot Setting

In this section, we introduce conditional Rényi entropies for the single-shot setting. For more a detailed review of conditional Rényi entropies, see [21]. For a correlated random variable (X, Y) on $\mathcal{X} \times \mathcal{Y}$ with probability distribution P_{XY} and a marginal distribution Q_Y on $\mathcal{Y}$, we introduce the conditional Rényi entropy of order $1+\theta$ relative to Q_Y as:

$$H_{1+\theta}(P_{XY}|Q_Y) := -\frac{1}{\theta}\log\sum_{x,y} P_{XY}(x,y)^{1+\theta} Q_Y(y)^{-\theta}, \tag{6}$$

where $\theta \in (-1, 0) \cup (0, \infty)$. The conditional Rényi entropy of order zero relative to Q_Y is defined by the limit with respect to θ. When X has no side-information, it is nothing but the ordinary Rényi entropy, and it is denoted by $H_{1+\theta}(X) = H_{1+\theta}(P_X)$ throughout the paper.

One of the important special cases of $H_{1+\theta}(P_{XY}|Q_Y)$ is the case with $Q_Y = P_Y$, where P_Y is the marginal of P_{XY}. We shall call this special case the lower conditional Rényi entropy of order $1+\theta$ and denote (this notation was first introduced in [56]):

$$\begin{aligned} H^{\downarrow}_{1+\theta}(X|Y) &:= H_{1+\theta}(P_{XY}|P_Y) && (7)\\ &= -\frac{1}{\theta}\log\sum_{x,y} P_{XY}(x,y)^{1+\theta}P_Y(y)^{-\theta}. && (8)\end{aligned}$$

When we consider the second-order analysis, the variance of the entropy density plays an important role:

$$\mathsf{V}(X|Y) := \mathrm{Var}\left[\log\frac{1}{P_{X|Y}(X|Y)}\right]. \tag{9}$$

We have the following property, which follows from the correspondence between the conditional Rényi entropy and the cumulant generating function (cf. Appendix B).

Lemma 1. *We have:*

$$\lim_{\theta\to 0} H^{\downarrow}_{1+\theta}(X|Y) = H(X|Y) \tag{10}$$

and (as seen in the proof (cf. (A26)*), the left-hand side of* (11) *corresponds to the second derivative of the cumulant generating function):*

$$\lim_{\theta\to 0}\frac{2\left[H(X|Y) - H^{\downarrow}_{1+\theta}(X|Y)\right]}{\theta} = \mathsf{V}(X|Y). \tag{11}$$

Proof. (10) follows from the relation in (A25) and the fact that the first-order derivative of the cumulant generating function is the expectation. (11) follows from (A25), (10) and (A26). □

The other important special case of $H_{1+\theta}(P_{XY}|Q_Y)$ is the measure maximized over Q_Y. We shall call this special case the upper conditional Rényi entropy of order $1+\theta$ and denote (Equation (13) for $-1 < \theta < 0$ follows from the Hölder inequality, and Equation (13) for $0 < \theta$ follows from the reverse Hölder inequality [57] (Lemma 8). Similar optimization has appeared in the context of Rényi mutual information in [58] (see also [59]).):

$$\begin{aligned} H^{\uparrow}_{1+\theta}(X|Y) &:= \max_{Q_Y\in\mathcal{P}(\mathcal{Y})} H_{1+\theta}(P_{XY}|Q_Y) && (12)\\ &= H_{1+\theta}(P_{XY}|P_Y^{(1+\theta)}) && (13)\\ &= -\frac{1+\theta}{\theta}\log\sum_y P_Y(y)\left[\sum_x P_{X|Y}(x|y)^{1+\theta}\right]^{\frac{1}{1+\theta}}, && (14)\end{aligned}$$

where:

$$P_Y^{(1+\theta)}(y) := \frac{\left[\sum_x P_{XY}(x,y)^{1+\theta}\right]^{\frac{1}{1+\theta}}}{\sum_{y'}\left[\sum_x P_{XY}(x,y')^{1+\theta}\right]^{\frac{1}{1+\theta}}}. \tag{15}$$

For this measure, we also have the same properties as Lemma 1. This lemma will be proven in Appendix C.

Lemma 2. *We have:*

$$\lim_{\theta \to 0} H^{\uparrow}_{1+\theta}(X|Y) = H(X|Y) \tag{16}$$

and:

$$\lim_{\theta \to 0} \frac{2\left[H(X|Y) - H^{\uparrow}_{1+\theta}(X|Y)\right]}{\theta} = \mathsf{V}(X|Y). \tag{17}$$

When we derive converse bounds, we need to consider the case such that the order of the Rényi entropy is different from the order of conditioning distribution defined in (15). For this purpose, we introduce two-parameter conditional Rényi entropy, which connects the two kinds of conditional Rényi entropies $H^{\downarrow}_{1+\theta}(X|Y)$ and $H^{\uparrow}_{1+\theta}(X|Y)$ in the way as Statements 10 and 11 of Lemma 3:

$$\begin{aligned} & H_{1+\theta,1+\theta'}(X|Y) && (18)\\ & := \quad H_{1+\theta}(P_{XY}|P_Y^{(1+\theta')}) && (19)\\ & = \quad -\frac{1}{\theta}\log \sum_y P_Y(y)\left[\sum_x P_{X|Y}(x|y)^{1+\theta}\right]\left[\sum_x P_{X|Y}(x|y)^{1+\theta'}\right]^{\frac{-\theta}{1+\theta'}} + \frac{\theta'}{1+\theta'} H^{\uparrow}_{1+\theta'}(X|Y). && (20) \end{aligned}$$

Next, we investigate some properties of the measures defined above, which will be proven in Appendix D.

Lemma 3.

1. *For fixed Q_Y, $\theta H_{1+\theta}(P_{XY}|Q_Y)$ is a concave function of θ, and it is strict concave iff* $\mathrm{Var}\left[\log \frac{Q_Y(Y)}{P_{XY}(X,Y)}\right] > 0$.
2. *For fixed Q_Y, $H_{1+\theta}(P_{XY}|Q_Y)$ is a monotonically decreasing (Technically, $H_{1+\theta}(P_{XY}|Q_Y)$ is always non-increasing, and it is monotonically decreasing iff strict concavity holds in Statement 1. Similar remarks are also applied for other information measures throughout the paper.) function of θ.*
3. *The function $\theta H^{\downarrow}_{1+\theta}(X|Y)$ is a concave function of θ, and it is strict concave iff $\mathsf{V}(X|Y) > 0$.*
4. *$H^{\downarrow}_{1+\theta}(X|Y)$ is a monotonically decreasing function of θ.*
5. *The function $\theta H^{\uparrow}_{1+\theta}(X|Y)$ is a concave function of θ, and it is strict concave iff $\mathsf{V}(X|Y) > 0$.*
6. *$H^{\uparrow}_{1+\theta}(X|Y)$ is a monotonically decreasing function of θ.*
7. *For every $\theta \in (-1,0) \cup (0,\infty)$, we have $H^{\downarrow}_{1+\theta}(X|Y) \le H^{\uparrow}_{1+\theta}(X|Y)$.*
8. *For fixed θ', the function $\theta H_{1+\theta,1+\theta'}(X|Y)$ is a concave function of θ, and it is strict concave iff $\mathsf{V}(X|Y) > 0$.*
9. *For fixed θ', $H_{1+\theta,1+\theta'}(X|Y)$ is a monotonically decreasing function of θ.*
10. *We have:*

$$H_{1+\theta,1}(X|Y) = H^{\downarrow}_{1+\theta}(X|Y). \tag{21}$$

11. *We have:*

$$H_{1+\theta,1+\theta}(X|Y) = H^{\uparrow}_{1+\theta}(X|Y). \tag{22}$$

12. *For every $\theta \in (-1,0) \cup (0,\infty)$, $H_{1+\theta,1+\theta'}(X|Y)$ is maximized at $\theta' = \theta$.*

The following lemma expresses explicit forms of the conditional Rényi entropies of order zero.

Lemma 4. *We have:*

$$\begin{aligned}
\lim_{\theta\to-1} H_{1+\theta}(P_{XY}|Q_Y) &= H_0(P_{XY}|Q_Y) && (23)\\
&:= \log\sum_y Q_Y(y)|\mathrm{supp}(P_{X|Y}(\cdot|y))|, && (24)\\
\lim_{\theta\to-1} H^{\uparrow}_{1+\theta}(X|Y) &= H^{\uparrow}_0(X|Y) && (25)\\
&:= \log\max_{y\in\mathrm{supp}(P_Y)} |\mathrm{supp}(P_{X|Y}(\cdot|y))|, && (26)\\
\lim_{\theta\to-1} H^{\downarrow}_{1+\theta}(X|Y) &= H^{\downarrow}_0(X|Y) && (27)\\
&:= \log\sum_y P_Y(y)|\mathrm{supp}(P_{X|Y}(\cdot|y))|. && (28)
\end{aligned}$$

Proof. See Appendix E. □

The definition (6) guarantees the existence of the derivative of $\frac{d[\theta H_{1+\theta}(P_{XY}|Q_Y)]}{d\theta}$. From Statement 1 of Lemma 3, $d[\theta H_{1+\theta}(P_{XY}|Q_Y)]/d\theta$ is monotonically decreasing. Thus, the inverse function (Throughout the paper, the notations $\theta(a)$ and $a(R)$ are reused for several inverse functions. Although the meanings of those notations are obvious from the context, we occasionally put superscript Q, $\downarrow$ or $\uparrow$ to emphasize that those inverse functions are induced from corresponding conditional Rényi entropies. This definition is related to the Legendre transform of the concave function $\theta \mapsto \theta H^{\downarrow}_{1+\theta}(X|Y)$.) of $\theta \mapsto d[\theta H_{1+\theta}(P_{XY}|Q_Y)]/d\theta$ exists so that the function $\theta(a) = \theta^Q(a)$ is defined as:

$$\left.\frac{d[\theta H_{1+\theta}(P_{XY}|Q_Y)]}{d\theta}\right|_{\theta=\theta(a)} = a \tag{29}$$

for $\underline{a} < a \le \overline{a}$, where $\underline{a} = \underline{a}^Q := \lim_{\theta\to\infty} d[\theta H_{1+\theta}(P_{XY}|Q_Y)]/d\theta$ and $\overline{a} = \overline{a}^Q := \lim_{\theta\to-1} d[\theta H_{1+\theta}(P_{XY}|Q_Y)]/d\theta$. Let:

$$R(a) = R^Q(a) := (1+\theta(a))a - \theta(a)H_{1+\theta(a)}(P_{XY}|Q_Y). \tag{30}$$

Since:

$$\begin{aligned}
R'(a) &= \frac{dR'(a)}{da}\\
&= \frac{d\theta(a)}{da}a + 1 + \theta(a) - \frac{d(\theta H_{1+\theta}(P_{XY}|Q_Y))}{d\theta}\frac{d\theta(a)}{da}\\
&= \frac{d\theta(a)}{da}a + 1 + \theta(a) - a\frac{d\theta(a)}{da} = 1+\theta(a),
\end{aligned} \tag{31}$$

$R(a)$ is a monotonic increasing function of $\underline{a} < a \le R(\overline{a})$. Thus, we can define the inverse function $a(R) = a^Q(R)$ of $R(a)$ by:

$$(1+\theta(a(R)))a(R) - \theta(a(R))H_{1+\theta(a(R))}(P_{XY}|Q_Y) = R \tag{32}$$

for $R(\underline{a}) < R \le H_0(P_{XY}|Q_Y)$.

For $\theta H^{\downarrow}_{1+\theta}(X|Y)$, by the same reason as above, we can define the inverse functions $\theta(a) = \theta^{\downarrow}(a)$ and $a(R) = a^{\downarrow}(R)$ by:

$$\left.\frac{d[\theta H^{\downarrow}_{1+\theta}(X|Y)]}{d\theta}\right|_{\theta=\theta(a)} = a \tag{33}$$

and:

$$(1+\theta(a(R)))a(R)-\theta(a(R))H^{\downarrow}_{1+\theta(a(R))}(X|Y)=R, \tag{34}$$

for $R(\underline{a})<R\le H^{\downarrow}_{0}(X|Y)$. For $\theta H^{\uparrow}_{1+\theta}(X|Y)$, we also introduce the inverse functions $\theta(a)=\theta^{\uparrow}(a)$ and $a(R)=a^{\uparrow}(R)$ by:

$$\left.\frac{d\theta H^{\uparrow}_{1+\theta}(X|Y)}{d\theta}\right|_{\theta=\theta(a)}=a \tag{35}$$

and:

$$(1+\theta(a(R)))a(R)-\theta(a(R))H^{\uparrow}_{1+\theta(a(R))}(X|Y)=R \tag{36}$$

for $R(\underline{a})<R\le H^{\uparrow}_{0}(X|Y)$.

Remark 1. *Here, we discuss the possibility for extension to the continuous case. Since the entropy in the continuous case diverges, we cannot extend the information quantities to the case when $\mathcal{X}$ is continuous. However, it is possible to extend these quantities to the case when $\mathcal{Y}$ is continuous, but $\mathcal{X}$ is a discrete finite set. In this case, we prepare a general measure μ (like the Lebesgue measure) on $\mathcal{Y}$ and probability density function p_Y and q_Y such that the distributions P_Y and Q_Y are given as $p_Y(y)\mu(dy)$ and $q_Y(y)\mu(dy)$, respectively. Then, it is sufficient to replace $\sum$, $Q(y)$, and $P_{XY}(x,y)$ by $\int_{\mathcal{Y}}\mu(dy)$, $P_{X|Y}(x|y)p_Y(y)$, and $q_Y(y)$, respectively. Hence, in the n-independent and identically distributed case, these information measures are given as n times the original information measures.*

One might consider the information quantities for transition matrices given in the next subsection for this continuous case. However, this is not so easy because it needs a continuous extension of the Perron eigenvalue.

2.2. Information Measures for the Transition Matrix

Let $\{W(x,y|x',y')\}_{((x,y),(x',y'))\in(\mathcal{X}\times\mathcal{Y})^2}$ be an ergodic and irreducible transition matrix. The purpose of this section is to introduce transition matrix counterparts of those measures in Section 2.1. For this purpose, we first need to introduce some assumptions on transition matrices:

Assumption 1 (Non-hidden)**.** *We say that a transition matrix W is non-hidden (with respect to Y) if the Y-marginal process is a Markov process, i.e., (The reason for the name "non-hidden" is the following. In general, the Y-marginal process is a hidden Markov process. However, when the condition (37) holds, the Y-marginal process is a Markov process. Hence, we call the condition (37) non-hidden.):*

$$\sum_{x}W(x,y|x',y')=W(y|y') \tag{37}$$

for every $x'\in\mathcal{X}$ and $y,y'\in\mathcal{Y}$. This condition is equivalent to the existence of the following decomposition of $W(x,y|x',y')$:

$$W(x,y|x',y')=W(y|y')W(x|x',y',y). \tag{38}$$

Assumption 2 (Strongly non-hidden)**.** *We say that a transition matrix W is strongly non-hidden (with respect to Y) if, for every $\theta\in(-1,\infty)$ and $y,y'\in\mathcal{Y}$ (The reason for the name "strongly non-hidden" is the following. When we compute the upper conditional Rényi entropy rate of the Markov source, the effect of the*

Y process may propagate infinitely even if it is non-hidden. When (39) holds, the effect of the Y process in the computation of the upper conditional Rényi entropy rate is only one step.):

$$W_\theta(y|y') := \sum_x W(x,y|x',y')^{1+\theta} \tag{39}$$

is well defined, i.e., the right-hand side of (39) is independent of x'.

Assumption 1 requires (39) to hold only for $\theta = 0$, and thus, Assumption 2 implies Assumption 1. However, Assumption 2 is a strictly stronger condition than Assumption 1. For example, let us consider the case such that the transition matrix is a product form, i.e., $W(x,y|x',y') = W(x|x')W(y|y')$. In this case, Assumption 1 is obviously satisfied. However, Assumption 2 is not satisfied in general.

Assumption 2 has another expression as follows.

Lemma 5. *Assumption 2 holds if and only if, for every $x' \neq \tilde{x}'$, there exists a permutation $\pi_{x';\tilde{x}'}$ on $\mathcal{X}$ such that $W(x|x',y',y) = W(\pi_{x';\tilde{x}'}(x)|\tilde{x}',y',y)$.*

Proof. Since the part "if" is trivial, we show the part "only if" as follows. By noting (38), Assumption 2 can be rephrased as:

$$\sum_x W(x|x',y',y)^{1+\theta} \tag{40}$$

does not depend on x' for every $\theta \in (-1,\infty)$. Furthermore, this condition can be rephrased as follows. For $x' \neq \tilde{x}'$, if the largest values of $\{W(x|x',y')\}_{x\in\mathcal{X}}$ and $\{W(x|\tilde{x}',y')\}_{x\in\mathcal{X}}$ are different, say the former is larger, then $\sum_x W(x|x',y')^{1+\theta} > \sum_x W(x|\tilde{x}',y')^{1+\theta}$ for sufficiently large θ, which contradicts the fact that (40) does not depend on x'. Thus, the largest values of $\{W(x|x',y')\}_{x\in\mathcal{X}}$ and $\{W(x|\tilde{x}',y')\}_{x\in\mathcal{X}}$ must coincide. By repeating this argument for the second largest value of $\{W(x|x',y')\}_{x\in\mathcal{X}}$ and $\{W(x|\tilde{x}',y')\}_{x\in\mathcal{X}}$, and so on, we find that Assumption 2 implies that for every $x' \neq \tilde{x}'$, there exists a permutation$\pi_{x';\tilde{x}'}$ on $\mathcal{X}$ such that $W(x|x',y',y) = W(\pi_{x';\tilde{x}'}(x)|\tilde{x}',y',y)$. □

Now, we fix an element $x_0 \in \mathcal{X}$ and transform a sequence of random numbers $(X_1,Y_1,X_2,Y_2,\ldots,X_n,Y_n)$ to the sequence of random numbers $(X'_1,Y'_1,X'_2,Y'_2,\ldots,X'_n,Y'_n) := (X_1,Y_1,\pi^{-1}_{x_0;X_1}(X_2),Y_2,\ldots,\pi^{-1}_{x_0;X_1}(X_n),Y_n)$. Then, letting $W'(x|y',y) := W(x|x_0,y',y)$, we have $P_{X'_i,Y'_i|X'_{i-1},Y'_{i-1}} = W'(y'_i|y'_{i-1})W(x'_i|y'_i,y'_{i-1})$. That is, essentially, the transition matrix of this case can be written by the transition matrix $W(y'_i|y'_{i-1})W'(x'_i|y'_i,y'_{i-1})$. Therefore, the transition matrix can be written by using the positive-entry matrix $W_{x'_i}(y'_i|y'_{i-1}) := W(y'_i|y'_{i-1})W'(x'_i|y'_i,y'_{i-1})$.

The following are non-trivial examples satisfying Assumptions 1 and 2.

Example 1. *Suppose that $\mathcal{X} = \mathcal{Y}$ is a module (an additive group). Let P and Q be transition matrices on $\mathcal{X}$. Then, the transition matrix given by:*

$$W(x,y|x',y') = Q(y|y')P(x-y|x'-y') \tag{41}$$

satisfies Assumption 1. Furthermore, if transition matrix $P(z|z')$ can be written as:

$$P(z|z') = P_Z(\pi_{z'}(z)) \tag{42}$$

for permutation $\pi_{z'}$ and a distribution P_Z on $\mathcal{X}$, then transition matrix W defined by (41) satisfies Assumption 2 as well.

Example 2. *Suppose that $\mathcal{X}$ is a module and W is (strongly) non-hidden with respect to $\mathcal{Y}$. Let Q be a transition matrix on $\mathcal{Z} = \mathcal{X}$. Then, the transition matrix given by:*

$$V(x, y, z|x', y', z') = W(x - z, y|x' - z', y)Q(z|z') \tag{43}$$

is (strongly) non-hidden with respect to $\mathcal{Y} \times \mathcal{Z}$.

The following is also an example satisfying Assumption 2, which describes a noise process of an important class of channels with memory (cf. the Gilbert-Elliot channel in Example 6).

Example 3. *Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. Then, let:*

$$W(y|y') = \begin{cases} 1 - q_{y'} & \text{if } y = y' \\ q_{y'} & \text{if } y \neq y' \end{cases} \tag{44}$$

for some $0 < q_0, q_1 < 1$, and let:

$$W(x|x', y', y) = \begin{cases} 1 - p_y & \text{if } x = 0 \\ p_y & \text{if } x = 1 \end{cases} \tag{45}$$

for some $0 < p_0, p_1 < 1$. By choosing $\pi_{x';\tilde{x}'}$ to be the identity, this transition matrix satisfies the condition given in Remark 5, which is equivalent to Assumption 2.

First, we introduce information measures under Assumption 1. In order to define a transition matrix counterpart of (7), let us introduce the following tilted matrix:

$$\tilde{W}_\theta(x, y|x', y') := W(x, y|x', y')^{1+\theta} W(y|y')^{-\theta}. \tag{46}$$

Here, we should notice that the tilted matrix $\tilde{W}_\theta$ is not normalized, i.e., is not a transition matrix. Let λ_θ be the Perron–Frobenius eigenvalue of $\tilde{W}_\theta$ and $\tilde{P}_{\theta,XY}$ be its normalized eigenvector. Then, we define the lower conditional Rényi entropy for W by:

$$H_{1+\theta}^{\downarrow,W}(X|Y) := -\frac{1}{\theta} \log \lambda_\theta, \tag{47}$$

where $\theta \in (-1, 0) \cup (0, \infty)$. For $\theta = 0$, we define the lower conditional Rényi entropy for W by:

$$H^W(X|Y) = H_1^{\downarrow,W}(X|Y) \tag{48}$$

$$:= \lim_{\theta \to 0} H_{1+\theta}^{\downarrow,W}(X|Y), \tag{49}$$

and we just call it the conditional entropy for W. In fact, the definition of $H^W(X|Y)$ above coincides with:

$$-\sum_{x',y'} P_{0,XY}(x', y') \sum_{x,y} W(x, y|x', y') \log \frac{W(x, y|x', y')}{W(y|y')}, \tag{50}$$

where $P_{0,XY}$ is the stationary distribution of W (cf. [60] (Equation (30))). For $\theta = -1$, $H_0^{\downarrow,W}(X|Y)$ is also defined by taking the limit. When X has no side-information, the Rényi entropy $H_{1+\theta}^W(X)$ for W is defined as a special case of $H_{1+\theta}^{\downarrow,W}(X|Y)$.

As a counterpart of (11), we also define (Since the limiting expression in (51) coincides with the second derivative of the CGF (cf. (A30)) and since the second derivative of the CGF exists (cf. [22] (Appendix D)), the variance in (51) is well defined. While the definition (51) contains the limit $\theta \to 0$, it

can be calculated without this type of limit by using the fundamental matrix [61] (Theorem 4.3.1), [23] (Theorem 7.7 and Remark 7.8).):

$$\mathsf{V}^W(X|Y) := \lim_{\theta \to 0} \frac{2\left[H^W(X|Y) - H^{\downarrow,W}_{1+\theta}(X|Y)\right]}{\theta}. \tag{51}$$

Remark 2. *When transition matrix W satisfies Assumption 2, $H^{\downarrow,W}_{1+\theta}(X|Y)$ can be written as:*

$$H^{\downarrow,W}_{1+\theta}(X|Y) = -\frac{1}{\theta} \log \lambda'_\theta, \tag{52}$$

where λ'_θ is the Perron–Frobenius eigenvalue of $W_\theta(y|y')W(y|y')^{-\theta}$. In fact, for the left Perron–Frobenius eigenvector $\hat{Q}_\theta$ of $W_\theta(y|y')W(y|y')^{-\theta}$, we have:

$$\sum_{x,y} \hat{Q}_\theta(y) W(x,y|x',y')^{1+\theta} W(y|y')^{-\theta} = \lambda'_\theta Q_\theta(y'), \tag{53}$$

which implies that λ'_θ is the Perron–Frobenius eigenvalue of $\tilde{W}_\theta$. Consequently, we can evaluate $H^{\downarrow,W}_{1+\theta}(X|Y)$ by calculating the Perron–Frobenius eigenvalue of the $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix instead of the $|\mathcal{X}||\mathcal{Y}| \times |\mathcal{X}||\mathcal{Y}|$ matrix when W satisfies Assumption 2.

Next, we introduce information measures under Assumption 2. In order to define a transition matrix counterpart of (12), let us introduce the following $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix:

$$K_\theta(y|y') := W_\theta(y|y')^{\frac{1}{1+\theta}}, \tag{54}$$

where W_θ is defined by (39). Let κ_θ be the Perron–Frobenius eigenvalue of K_θ. Then, we define the upper conditional Rényi entropy for W by:

$$H^{\uparrow,W}_{1+\theta}(X|Y) := -\frac{1+\theta}{\theta} \log \kappa_\theta, \tag{55}$$

where $\theta \in (-1, 0) \cup (0, \infty)$. For $\theta = -1$ and $\theta = 0$, $H^{\uparrow,W}_{1+\theta}(X|Y)$ is defined by taking the limit. We have the following properties, which will be proven in Appendix F.

Lemma 6. *We have:*

$$\lim_{\theta \to 0} H^{\uparrow,W}_{1+\theta}(X|Y) = H^W(X|Y) \tag{56}$$

and:

$$\lim_{\theta \to 0} \frac{2\left[H^W(X|Y) - H^{\uparrow,W}_{1+\theta}(X|Y)\right]}{\theta} = \mathsf{V}^W(X|Y). \tag{57}$$

Now, let us introduce a transition matrix counterpart of (18). For this purpose, we introduce the following $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix:

$$N_{\theta,\theta'}(y|y') := W_\theta(y|y') W_{\theta'}(y|y')^{\frac{-\theta}{1+\theta'}}. \tag{58}$$

Let $\nu_{\theta,\theta'}$ be the Perron–Frobenius eigenvalue of $N_{\theta,\theta'}$. Then, we define the two-parameter conditional Rényi entropy by:

$$H^W_{1+\theta,1+\theta'}(X|Y) := -\frac{1}{\theta} \log \nu_{\theta,\theta'} + \frac{\theta'}{1+\theta'} H^{\uparrow,W}_{1+\theta'}(X|Y). \tag{59}$$

Remark 3. *Although we defined* $H_{1+\theta}^{\downarrow,W}(X|Y)$ *and* $H_{1+\theta}^{\uparrow,W}(X|Y)$ *by* (47) *and* (55)*, respectively, we can alternatively define these measures in the same spirit as the single-shot setting by introducing a transition matrix counterpart of* $H_{1+\theta}(P_{XY}|Q_Y)$ *as follows. For the marginal* $W(y|y')$ *of* $W(x,y|x',y')$*, let* $\mathcal{Y}_W^2 := \{(y,y') : W(y|y') > 0\}$*. For another transition matrix* V *on* $\mathcal{Y}$*, we define* $\mathcal{Y}_V^2$ *in a similar manner. For* V *satisfying* $\mathcal{Y}_W^2 \subset \mathcal{Y}_V^2$*, we define (although we can also define* $H_{1+\theta}^{W|V}(X|Y)$ *even if* $\mathcal{Y}_W^2 \subset \mathcal{Y}_V^2$ *is not satisfied (see [22] for the detail), for our purpose of defining* $H_{1+\theta}^{\downarrow,W}(X|Y)$ *and* $H_{1+\theta}^{\uparrow,W}(X|Y)$*, other cases are irrelevant):*

$$H_{1+\theta}^{W|V}(X|Y) := -\frac{1}{\theta}\log \lambda_\theta^{W|V} \tag{60}$$

for $\theta \in (-1,0)\cup(0,\infty)$*, where* $\lambda_\theta^{W|V}$ *is the Perron–Frobenius eigenvalue of:*

$$W(x,y|x',y')^{1+\theta}V(y|y')^{-\theta}. \tag{61}$$

By using this measure, we obviously have:

$$H_{1+\theta}^{\downarrow,W}(X|Y) = H_{1+\theta}^{W|W}(X|Y). \tag{62}$$

Furthermore, under Assumption 2, the relation:

$$H_{1+\theta}^{\uparrow,W}(X|Y) = \max_V H_{1+\theta}^{W|V}(X|Y) \tag{63}$$

holds (see Appendix G for the proof), where the maximum is taken over all transition matrices satisfying $\mathcal{Y}_W^2 \subset \mathcal{Y}_V^2$*.*

Next, we investigate some properties of the information measures introduced in this section. The following lemma is proven in Appendix H.

Lemma 7.

1. *The function* $\theta H_{1+\theta}^{\downarrow,W}(X|Y)$ *is a concave function of* θ*, and it is strict concave iff* $\mathsf{V}^W(X|Y) > 0$*.*
2. $H_{1+\theta}^{\downarrow,W}(X|Y)$ *is a monotonically decreasing function of* θ*.*
3. *The function* $\theta H_{1+\theta}^{\uparrow,W}(X|Y)$ *is a concave function of* θ*, and it is strict concave iff* $\mathsf{V}^W(X|Y) > 0$*.*
4. $H_{1+\theta}^{\uparrow,W}(X|Y)$ *is a monotonically decreasing function of* θ*.*
5. *For every* $\theta \in (-1,0)\cup(0,\infty)$*, we have* $H_{1+\theta}^{\downarrow,W}(X|Y) \le H_{1+\theta}^{\uparrow,W}(X|Y)$*.*
6. *For fixed* θ'*, the function* $\theta H_{1+\theta,1+\theta'}^{W}(X|Y)$ *is a concave function of* θ*, and it is strict concave iff* $\mathsf{V}^W(X|Y) > 0$*.*
7. *For fixed* θ'*,* $H_{1+\theta,1+\theta'}^{W}(X|Y)$ *is a monotonically decreasing function of* θ*.*
8. *We have:*

$$H_{1+\theta,1}^{W}(X|Y) = H_{1+\theta}^{\downarrow,W}(X|Y). \tag{64}$$

9. *We have:*

$$H_{1+\theta,1+\theta}^{W}(X|Y) = H_{1+\theta}^{\uparrow,W}(X|Y). \tag{65}$$

10. *For every* $\theta \in (-1,0)\cup(0,\infty)$*,* $H_{1+\theta,1+\theta'}^{W}(X|Y)$ *is maximized at* $\theta' = \theta$*, i.e.,*

$$\left.\frac{d[H_{1+\theta,1+\theta'}^{W}(X|Y)]}{d\theta'}\right|_{\theta'=\theta} = 0. \tag{66}$$

From Statement 1 of Lemma 7, $d[\theta H^{\downarrow,W}_{1+\theta}(X|Y)]/d\theta$ is monotonically decreasing. Thus, we can define the inverse function $\theta^W(a) = \theta^{\downarrow,W}(a)$ of $d[\theta H^{\downarrow,W}_{1+\theta}(X|Y)]/d\theta$ by:

$$\left.\frac{d[\theta H^{\downarrow,W}_{1+\theta}(X|Y)]}{d\theta}\right|_{\theta=\theta^W(a)} = a \tag{67}$$

for $\underline{a} < a \leq \overline{a}$, where $\underline{a} := \lim_{\theta\to\infty} d[\theta H^{\downarrow,W}_{1+\theta}(X|Y)]/d\theta$ and $\overline{a} := \lim_{\theta\to-1} d[\theta H^{\downarrow,W}_{1+\theta}(X|Y)]/d\theta$. Let:

$$R^W(a) := (1+\theta(a))a - \theta(a)H^{\downarrow,W}_{1+\theta(a)}(X|Y). \tag{68}$$

Since

$$R^{W\prime}(a) = (1+\theta(a)), \tag{69}$$

$R^W(a)$ is a monotonic increasing function of $\underline{a} < a < R^W(\overline{a})$. Thus, we can define the inverse function $a^W(R) = a^{\downarrow,W}(R)$ of $R^W(a)$ by:

$$(1+\theta(a^W(R)))a^W(R) - \theta^W(a^W(R))H^{\downarrow,W}_{1+\theta^W(a^W(R))}(X|Y) = R \tag{70}$$

for $R^W(\underline{a}) < R < H^{\downarrow,W}_0(X|Y)$, where $H^{\downarrow,W}_0(X|Y) := \lim_{\theta\to-1} H^{\downarrow,W}_{1+\theta}(X|Y)$.

For $\theta H^{\uparrow,W}_{1+\theta}(X|Y)$, by the same reason, we can define the inverse function $\theta^W(a) = \theta^{\uparrow,W}(a)$ by:

$$\left.\frac{d[\theta H^{W}_{1+\theta,1+\theta^W(a)}(X|Y)]}{d\theta}\right|_{\theta=\theta^W(a)} = \left.\frac{d[\theta H^{\uparrow,W}_{1+\theta}(X|Y)]}{d\theta}\right|_{\theta=\theta^W(a)} = a, \tag{71}$$

and the inverse function $a^W(R) = a^{\uparrow,W}(R)$ of:

$$R^W(a) := (1+\theta^W(a))a - \theta^W(a)H^{\uparrow,W}_{1+\theta^W(a)}(X|Y) \tag{72}$$

by:

$$(1+\theta^W(a^W(R)))a^W(R) - \theta^W(a^W(R))H^{\uparrow,W}_{1+\theta^W(a^W(R))}(X|Y) = R, \tag{73}$$

for $R(\underline{a}) < R < H^{\uparrow,W}_0(X|Y)$, where $H^{\uparrow,W}_0(X|Y) := \lim_{\theta\to-1} H^{\uparrow,W}_{1+\theta}(X|Y)$. Here, the first equality in (71) follows from (66).

Since $\theta \mapsto \theta H^{\downarrow,W}_{1+\theta}(X|Y)$ is concave, the supremum of $[-\theta R + \theta H^{\downarrow,W}_{1+\theta}(X|Y)]$ is attained at the stationary point. Furthermore, note that $-1 \leq \theta^{\downarrow,W}(R) \leq 0$ for $H^W(X|Y) \leq R \leq H^{\downarrow,W}_0(X|Y)$. Thus, we have the following property.

Lemma 8. *The function $\theta^W(R)$ defined in (67) satisfies:*

$$\sup_{-1\leq\theta\leq 0}[-\theta R + \theta H^{\downarrow,W}_{1+\theta}(X|Y)] = -\theta^W(R)R + \theta^W(R)H^{\downarrow,W}_{1+\theta^W(R)}(X|Y) \tag{74}$$

for $H^W(X|Y) \leq R \leq H^{\downarrow,W}_0(X|Y)$.

Furthermore, we have the following characterization for another type of maximization.

Lemma 9. *The function $\theta^W(a^W(R))$ defined by (70) satisfies:*

$$\sup_{-1\le\theta\le 0}\frac{-\theta R+\theta H_{1+\theta}^{\downarrow,W}(X|Y)}{1+\theta}=-\theta^W(a^W(R))a^W(R)+\theta^W(a^W(R))H_{1+\theta^W(a^W(R))}^{\downarrow,W}(X|Y) \tag{75}$$

for $H^W(X|Y)\le R\le H_0^{\downarrow,W}(X|Y)$, and the function $\theta(a(R))$ defined in (73) satisfies:

$$\sup_{-1\le\theta\le 0}\frac{-\theta R+\theta H_{1+\theta}^{\uparrow,W}(X|Y)}{1+\theta}=-\theta^W(a^W(R))a^W(R)+\theta^W(a^W(R))H_{1+\theta^W(a^W(R))}^{\uparrow,W}(X|Y) \tag{76}$$

for $H^W(X|Y)\le R\le H_0^{\uparrow,W}(X|Y)$.

Proof. See Appendix I. □

Remark 4. *The combination of (49), (51), and Lemma 6 guarantees that both the conditional Rényi entropies expand as:*

$$H_{1+\theta}^{\downarrow,W}(X|Y) = H^W(X|Y)-\frac{1}{2}\mathrm{V}^W(X|Y)\theta+o(\theta), \tag{77}$$

$$H_{1+\theta}^{\uparrow,W}(X|Y) = H^W(X|Y)-\frac{1}{2}\mathrm{V}^W(X|Y)\theta+o(\theta) \tag{78}$$

around $\theta=0$. Thus, the difference of these measures significantly appears only when $|\theta|$ is rather large. For the transition matrix of Example 3 with $q_0=q_1=0.1$, $p_0=0.1$, and $p_1=0.4$, we plotted the values of the information measures in Figure 1. Although the values at $\theta=-1$ coincide in Figure 1, note that the values at $\theta=-1$ may differ in general.

In Example 1, we mentioned that the transition matrix W in (41) satisfies Assumption 2 when transition matrix P is given by (42). By computing the conditional Rényi entropies for this special case, we have:

$$H_{1+\theta}^{\uparrow,W}(X|Y) = H_{1+\theta}^{\downarrow,W}(X|Y) \tag{79}$$

$$= H_{1+\theta}(P_Z), \tag{80}$$

i.e., the two kinds of conditional Rényi entropies coincide.

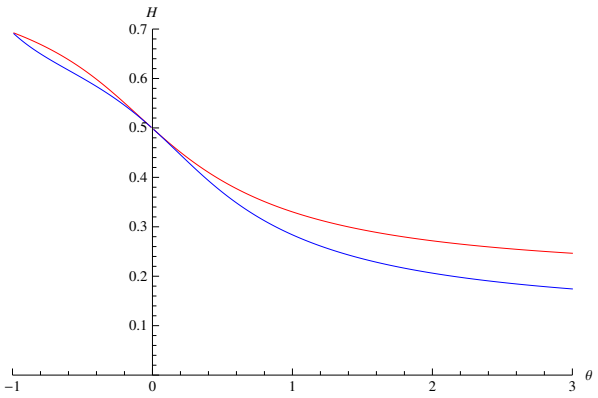

Figure 1. A comparison of $H_{1+\theta}^{\uparrow,W}(X|Y)$ (upper red curve) and $H_{1+\theta}^{\downarrow,W}(X|Y)$ (lower blue curve) for the transition matrix of Example 3 with $q_0=q_1=0.1$, $p_0=0.1$, and $p_1=0.4$. The horizontal axis is θ, and the vertical axis is the values of the information measures (nats).

Now, let us consider the asymptotic behavior of $H^{\downarrow,W}_{1+\theta}(X|Y)$ around $\theta = 0$. When $\theta(a)$ is close to zero, we have:

$$\theta^W(a)H^{\downarrow,W}_{1+\theta^W(a)}(X|Y) = \theta^W(a)H^W(X|Y) - \frac{1}{2}\mathsf{V}^W(X|Y)\theta^W(a)^2 + o(\theta^W(a)^2). \tag{81}$$

Taking the derivative, (67) implies that:

$$a = H^W(X|Y) - \mathsf{V}^W(X|Y)\theta^W(a) + o(\theta^W(a)). \tag{82}$$

Hence, when R is close to $H^W(X|Y)$, we have:

$$R = (1+\theta^W(a^W(R))a^W(R) - \theta^W(a^W(R))H^{\downarrow,W}_{1+\theta^W(a^W(R))}(X|Y) \tag{83}$$

$$= H^W(X|Y) - \left(1 + \frac{\theta^W(a^W(R))}{2}\right)\theta^W(a^W(R))\mathsf{V}^W(X|Y) + o(\theta^W(a^W(R)), \tag{84}$$

i.e.,

$$\theta^W(a^W(R)) = \frac{-R + H^W(X|Y)}{\mathsf{V}^W(X|Y)} + o\left(\frac{R - H^W(X|Y)}{\mathsf{V}^W(X|Y)}\right). \tag{85}$$

Furthermore, (81) and (82) imply:

$$-\theta^W(a^W(R))a^W(R) + \theta^W(a^W(R))H^{\downarrow,W}_{1+\theta^W(a^W(R))}(X|Y) \tag{86}$$

$$= \mathsf{V}^W(X|Y)\frac{\theta^W(a^W(R))^2}{2} + o(\theta^W(a^W(R))^2) \tag{87}$$

$$= \frac{\mathsf{V}^W(X|Y)}{2}\left(\frac{R - H^W(X|Y)}{\mathsf{V}^W(X|Y)}\right)^2 + o\left(\left(\frac{R - H^W(X|Y)}{\mathsf{V}^W(X|Y)}\right)^2\right). \tag{88}$$

2.3. Information Measures for the Markov Chain

Let $(\mathbf{X}, \mathbf{Y})$ be the Markov chain induced by transition matrix W and some initial distribution $P_{X_1Y_1}$. Now, we show how information measures introduced in Section 2.2 are related to the conditional Rényi entropy rates. First, we introduce the following lemma, which gives finite upper and lower bounds on the lower conditional Rényi entropy.

Lemma 10. *Suppose that transition matrix W satisfies Assumption 1. Let v_θ be the eigenvector of W^T_θ with respect to the Perron–Frobenius eigenvalue λ_θ such that $\min_{x,y} v_\theta(x,y) = 1$ (since the eigenvector corresponding to the Perron–Frobenius eigenvalue for an irreducible non-negative matrix has always strictly positive entries [62] (Theorem 8.4.4, p. 508), we can choose the eigenvector v_θ satisfying this condition). Let $w_\theta(x,y) := P_{X_1Y_1}(x,y)^{1+\theta}P_{Y_1}(y)^{-\theta}$. Then, for every $n \geq 1$, we have:*

$$(n-1)\theta H^{\downarrow,W}_{1+\theta}(X|Y) + \underline{\delta}(\theta) \leq \theta H^{\downarrow}_{1+\theta}(X^n|Y^n) \leq (n-1)\theta H^{\downarrow,W}_{1+\theta}(X|Y) + \overline{\delta}(\theta), \tag{89}$$

where:

$$\overline{\delta}(\theta) := -\log\langle v_\theta|w_\theta\rangle + \log\max_{x,y} v_\theta(x,y), \tag{90}$$

$$\underline{\delta}(\theta) := -\log\langle v_\theta|w_\theta\rangle, \tag{91}$$

and $\langle v_\theta|w_\theta\rangle$ is defined as $\sum_{x,y} v_\theta(x,y)w_\theta(x,y)$.

Proof. This follows from (A29) and Lemma A2. □

From Lemma 10, we have the following.

Theorem 1. *Suppose that transition matrix W satisfies Assumption 1. For any initial distribution, we have (When there is no side-information, (93) reduces to the well-known expression of the entropy rate of the Markov process [39]. Without Assumption 1, it is not clear if (93) holds or not.):*

$$\lim_{n\to\infty}\frac{1}{n}H^{\downarrow}_{1+\theta}(X^n|Y^n) = H^{\downarrow,W}_{1+\theta}(X|Y), \tag{92}$$

$$\lim_{n\to\infty}\frac{1}{n}H(X^n|Y^n) = H^{W}(X|Y). \tag{93}$$

We also have the following asymptotic evaluation of the variance, which follows from Lemma A3 in Appendix A.

Theorem 2. *Suppose that transition matrix W satisfies Assumption 1. For any initial distribution, we have:*

$$\lim_{n\to\infty}\frac{1}{n}\mathsf{V}(X^n|Y^n) = \mathsf{V}^{W}(X|Y). \tag{94}$$

Theorem 2 is practically important since the limit of the variance can be described by a single-letter characterized quantity. A method to calculate $\mathsf{V}^{W}(X|Y)$ can be found in [23].

Next, we show the lemma that gives the finite upper and lower bounds on the upper conditional Rényi entropy in terms of the upper conditional Rényi entropy for the transition matrix.

Lemma 11. *Suppose that transition matrix W satisfies Assumption 2. Let v_θ be the eigenvector of K_θ^T with respect to the Perron–Frobenius eigenvalue κ_θ such that $\min_y v_\theta(y) = 1$. Let w_θ be the $|\mathcal{Y}|$-dimensional vector defined by:*

$$w_\theta(y) := \left[\sum_x P_{X_1Y_1}(x,y)^{1+\theta}\right]^{\frac{1}{1+\theta}}. \tag{95}$$

Then, we have:

$$(n-1)\frac{\theta}{1+\theta}H^{\uparrow,W}_{1+\theta}(X|Y) + \underline{\xi}(\theta) \le \frac{\theta}{1+\theta}H^{\uparrow}_{1+\theta}(X^n|Y^n) \le (n-1)\frac{\theta}{1+\theta}H^{\uparrow,W}_{1+\theta}(X|Y) + \overline{\xi}(\theta), \tag{96}$$

where:

$$\overline{\xi}(\theta) := -\log\langle v_\theta|w_\theta\rangle + \log\max_y v_\theta(y), \tag{97}$$

$$\underline{\xi}(\theta) := -\log\langle v_\theta|w_\theta\rangle. \tag{98}$$

Proof. See Appendix J. □

From Lemma 11, we have the following.

Theorem 3. *Suppose that transition matrix W satisfies Assumption 2. For any initial distribution, we have:*

$$\lim_{n\to\infty}\frac{1}{n}H^{\uparrow}_{1+\theta}(X^n|Y^n) = H^{\uparrow,W}_{1+\theta}(X|Y). \tag{99}$$

Finally, we show the lemma that gives the finite upper and lower bounds on the two-parameter conditional Rényi entropy in terms of the two-parameter conditional Rényi entropy for the transition matrix.

Lemma 12. *Suppose that transition matrix W satisfies Assumption 2. Let $v_{\theta,\theta'}$ be the eigenvector of $N_{\theta,\theta'}^T$ with respect to the Perron–Frobenius eigenvalue $\nu_{\theta,\theta'}$ such that $\min_y v_{\theta,\theta'}(y) = 1$. Let $w_{\theta,\theta'}$ be the $|\mathcal{Y}|$-dimensional vector defined by:*

$$w_{\theta,\theta'}(y) := \left[\sum_x P_{X_1Y_1}(x,y)^{1+\theta}\right]\left[\sum_x P_{X_1Y_1}(x,y)^{1+\theta'}\right]^{\frac{-\theta}{1+\theta'}}. \tag{100}$$

Then, we have:

$$(n-1)\theta H^W_{1+\theta,1+\theta'}(X|Y) + \underline{\zeta}(\theta,\theta') \le \theta H_{1+\theta,1+\theta'}(X^n|Y^n) \le (n-1)\theta H^W_{1+\theta,1+\theta'}(X|Y) + \overline{\zeta}(\theta,\theta'), \tag{101}$$

where:

$$\overline{\zeta}(\theta,\theta') := -\log\langle v_{\theta,\theta'}|w_{\theta,\theta'}\rangle + \log\max_y v_{\theta,\theta'}(y) + \theta\overline{\xi}(\theta'), \tag{102}$$

$$\underline{\zeta}(\theta,\theta') := -\log\langle v_{\theta,\theta'}|w_{\theta,\theta'}\rangle + \theta\underline{\xi}(\theta') \tag{103}$$

for $\theta > 0$ and:

$$\overline{\zeta}(\theta,\theta') := -\log\langle v_{\theta,\theta'}|w_{\theta,\theta'}\rangle + \log\max_y v_{\theta,\theta'}(y) + \theta\underline{\xi}(\theta'), \tag{104}$$

$$\underline{\zeta}(\theta,\theta') := -\log\langle v_{\theta,\theta'}|w_{\theta,\theta'}\rangle + \theta\overline{\xi}(\theta') \tag{105}$$

for $\theta < 0$

Proof. By multiplying θ in the definition of $H_{1+\theta,1+\theta'}(X^n|Y^n)$, we have:

$$\theta H_{1+\theta,1+\theta'}(X^n|Y^n) \tag{106}$$

$$= -\log\sum_{y^n}\left[\sum_{x^n} P_{X^nY^n}(x^n,y^n)^{1+\theta}\right]\left[\sum_{x^n} P_{X^nY^n}(x^n,y^n)^{1+\theta'}\right]^{\frac{-\theta}{1+\theta'}} + \frac{\theta\theta'}{1+\theta'}H^{\uparrow}_{1+\theta'}(X^n|Y^n). \tag{107}$$

The second term is evaluated by Lemma 11. The first term can be evaluated almost in the same manner as Lemma 11. □

From Lemma 12, we have the following.

Theorem 4. *Suppose that transition matrix W satisfies Assumption 2. For any initial distribution, we have:*

$$\lim_{n\to\infty}\frac{1}{n}H_{1+\theta,1+\theta'}(X^n|Y^n) = H^W_{1+\theta,1+\theta'}(X|Y). \tag{108}$$

3. Source Coding with Full Side-Information

In this section, we investigate source coding with side-information. We start this section by showing the problem setting in Section 3.1. Then, we review and introduce some single-shot bounds in Section 3.2. We derive finite-length bounds for the Markov chain in Section 3.3. Then, in Sections 3.5 and 3.6, we show the asymptotic characterization for the large deviation regime and the moderate deviation regime by using those finite-length bounds. We also derive the second-order rate in Section 3.4.

3.1. Problem Formulation

A code $\Psi = (\mathsf{e}, \mathsf{d})$ consists of one encoder $\mathsf{e} : \mathcal{X} \to \{1, \ldots, M\}$ and one decoder $\mathsf{d} : \{1, \ldots, M\} \times \mathcal{Y} \to \mathcal{X}$. The decoding error probability is defined by:

$$\begin{aligned} \mathrm{P_s}[\Psi] &= \mathrm{P_s}[\Psi | P_{XY}] && (109)\\ &:= \Pr\{X \neq \mathsf{d}(\mathsf{e}(X), Y)\}. && (110) \end{aligned}$$

For notational convenience, we introduce the infimum of error probabilities under the condition that the message size is M:

$$\begin{aligned} \mathrm{P_s}(M) &= \mathrm{P_s}(M | P_{XY}) && (111)\\ &:= \inf_{\Psi} \mathrm{P_s}[\Psi]. && (112) \end{aligned}$$

For theoretical simplicity, we focus on a randomized choice of our encoder. For this purpose, we employ a randomized hash function F from $\mathcal{X}$ to $\{1, \ldots, M\}$. A randomized hash function F is called a two-universal hash when $\Pr\{F(x) = F(x')\} \leq \frac{1}{M}$ for any distinctive x and x' [63]; the so-called bin coding [39] is an example of the two-universal hash function. In the following, we denote the set of two-universal hash functions by $\mathcal{F}$. Given an encoder f as a function from $\mathcal{X}$ to $\{1, \ldots, M\}$, we define the decoder d_f as the optimal decoder by $\operatorname{argmin}_{\mathsf{d}} \mathrm{P_s}[(f, \mathsf{d})]$. Then, we denote the code (f, d_f) by $\Psi(f)$. Then, we bound the error probability $\mathrm{P_s}[\Psi(F)]$ averaged over the random function F by only using the property of two-universality. In order to consider the worst case of such schemes, we introduce the following quantity:

$$\begin{aligned} \bar{\mathrm{P}}_{\mathrm{s}}(M) &= \bar{\mathrm{P}}_{\mathrm{s}}(M | P_{XY}) && (113)\\ &:= \sup_{F \in \mathcal{F}} \mathbb{E}_F[\mathrm{P_s}[\Psi(F)]], . && (114) \end{aligned}$$

When we consider n-fold extension, the source code and related quantities are denoted with the superscript (n). For example, the quantities in (112) and (114) are written as $\mathrm{P}_{\mathrm{s}}^{(n)}(M)$ and $\bar{\mathrm{P}}_{\mathrm{s}}^{(n)}(M)$, respectively. Instead of evaluating them, we are often interested in evaluating:

$$\begin{aligned} M(n, \varepsilon) &:= \inf\{M_n : \mathrm{P}_{\mathrm{s}}^{(n)}(M_n) \leq \varepsilon\}, && (115)\\ \bar{M}(n, \varepsilon) &:= \inf\{M_n : \bar{\mathrm{P}}_{\mathrm{s}}^{(n)}(M_n) \leq \varepsilon\} && (116) \end{aligned}$$

for given $0 \leq \varepsilon < 1$.

3.2. Single-Shot Bounds

In this section, we review existing single-shot bounds and also show novel converse bounds. For the information measures used below, see Section 2.

By using the standard argument on information-spectrum approach, we have the following achievability bound.

Lemma 13 (Lemma 7.2.1 of [3]). *The following bound holds:*

$$\bar{\mathrm{P}}_{\mathrm{s}}(M) \leq \inf_{\gamma \geq 0} \left[P_{XY} \left\{ \log \frac{1}{P_{X|Y}(x|y)} > \gamma \right\} + \frac{e^{\gamma}}{M} \right]. \quad (117)$$

Although Lemma 13 is useful for the second-order regime, it is known to be not tight in the large deviation regime. By using the large deviation technique of Gallager, we have the following exponential-type achievability bound.

Lemma 14 ([64]). *The following bound holds: (note that the Gallager function and the upper conditional Rényi entropy are related by (A45)):*

$$\bar{\mathrm{P}}_{\mathrm{s}}(M) \leq \inf_{-\frac{1}{2}\leq\theta\leq 0} M^{\frac{\theta}{1+\theta}} e^{-\frac{\theta}{1+\theta} H^{\uparrow}_{1+\theta}(X|Y)}. \tag{118}$$

Although Lemma 14 is known to be tight in the large deviation regime for i.i.d. sources, $H^{\uparrow}_{1+\theta}(X|Y)$ for Markov chains can only be evaluated under the strongly non-hidden assumption. For this reason, even though the following bound is looser than Lemma 14, it is useful to have another bound in terms of $H^{\downarrow}_{1+\theta}(X|Y)$, which can be evaluated for Markov chains under the non-hidden assumption.

Lemma 15. *The following bound holds:*

$$\bar{\mathrm{P}}_{\mathrm{s}}(M) \leq \inf_{-1\leq\theta\leq 0} M^{\theta} e^{-\theta H^{\downarrow}_{1+\theta}(X|Y)}. \tag{119}$$

Proof. To derive this bound, we change the variable in (118) as $\theta = \frac{\theta'}{1-\theta'}$. Then, $-1 \leq \theta' \leq 0$, and we have:

$$M^{\theta'} e^{-\theta' H^{\uparrow}_{\frac{1}{1-\theta'}}(X|Y)} \leq M^{\theta'} e^{-\theta' H^{\downarrow}_{1+\theta'}(X|Y)},$$

where we use Lemma A4 in Appendix C. □

For the source coding without side-information, i.e., when X has no side-information, we have the following bound, which is tighter than Lemma 14.

Lemma 16 ((2.39) [65]). *The following bound holds:*

$$\mathrm{P}_{\mathrm{s}}(M) \leq \inf_{-1<\theta\leq 0} M^{\frac{\theta}{1+\theta}} e^{-\frac{\theta}{1+\theta} H_{1+\theta}(X)}. \tag{120}$$

For the converse part, we first have the following bound, which is very close to the operational definition of source coding with side-information.

Lemma 17 ([66]). *Let $\{\Omega_y\}_{y\in\mathcal{Y}}$ be a family of subsets $\Omega_y \subset \mathcal{X}$, and let $\Omega = \cup_{y\in\mathcal{Y}} \Omega_y \times \{y\}$. Then, for any $Q_Y \in \mathcal{P}(\mathcal{Y})$, the following bound holds:*

$$\mathrm{P}_{\mathrm{s}}(M) \geq \min_{\{\Omega_y\}} \left\{ P_{XY}(\Omega^c) : \sum_y Q_Y(y)|\Omega_y| \leq M \right\}. \tag{121}$$

Since Lemma 17 is close to the operational definition, it is not easy to evaluate Lemma 17. Thus, we derive another bound by loosening Lemma 17, which is more tractable for evaluation. Slightly weakening Lemma 17, we have the following.

Lemma 18 ([3,4]). *For any $Q_Y \in \mathcal{P}(\mathcal{Y})$, we have (In fact, a special case for $Q_Y = P_Y$ corresponds to Lemma 7.2.2 of [3]. A bound that involves Q_Y was introduced in [4] for channel coding, and it can be regarded as a source coding counterpart of that result.):*

$$\mathrm{P}_{\mathrm{s}}(M) \geq \sup_{\gamma\geq 0} \left[P_{XY}\left\{ \log \frac{Q_Y(y)}{P_{XY}(x,y)} > \gamma \right\} - \frac{M}{e^{\gamma}} \right]. \tag{122}$$

By using the change-of-measure argument, we also obtain the following converse bound.

Theorem 5. *For any $Q_Y \in \mathcal{P}(\mathcal{Y})$, we have:*

$$-\log \mathrm{P_s}(M) \tag{123}$$

$$\le \inf_{\substack{s>0 \\ \tilde{\theta}\in\mathbb{R},\vartheta\ge 0}} \Bigg[(1+s)\tilde{\theta}\left\{H_{1+\tilde{\theta}}(P_{XY}|Q_Y) - H_{1+(1+s)\tilde{\theta}}(P_{XY}|Q_Y)\right\} - (1+s)\log\left(1 - 2e^{-\frac{-\vartheta R + (\tilde{\theta}+\vartheta(1+\tilde{\theta}))H_{\tilde{\theta}+\vartheta(1+\tilde{\theta})}(P_{XY}|Q_Y) - (1+\vartheta)\tilde{\theta}H_{1+\tilde{\theta}}(P_{XY}|Q_Y)}{1+\vartheta}}\right)\Bigg]/s \tag{124}$$

$$\le \inf_{\substack{s>0 \\ -1<\tilde{\theta}<\theta(a(R))}} \Bigg[(1+s)\tilde{\theta}\left\{H_{1+\tilde{\theta}}(P_{XY}|Q_Y) - H_{1+(1+s)\tilde{\theta}}(P_{XY}|Q_Y)\right\} - (1+s)\log\left(1 - 2e^{(\theta(a(R))-\tilde{\theta})a(R) - \theta(a(R))H_{1+\theta(a(R))}(P_{XY}|Q_Y) + \tilde{\theta}H_{1+\tilde{\theta}}(P_{XY}|Q_Y)}\right)\Bigg]/s, \tag{125}$$

where $R = \log M$, and $\theta(a) = \theta^{\mathrm{Q}}(a)$ and $a(R) = a^{\mathrm{Q}}(R)$ are the inverse functions defined in (29) and (32), respectively.

Proof. See Appendix K. □

In particular, by taking $Q_Y = P_Y^{(1+\theta(a(R)))}$ in Theorem 5, we have the following.

Corollary 1. *We have:*

$$-\log \mathrm{P_s}(M) \tag{126}$$

$$\le \inf_{\substack{s>0 \\ -1<\tilde{\theta}<\theta(a(R))}} \Bigg[(1+s)\tilde{\theta}\left\{H_{1+\tilde{\theta},1+\theta(a(R))}(X|Y) - H_{1+(1+s)\tilde{\theta},1+\theta(a(R))}(X|Y)\right\} - (1+s)\log\left(1 - 2e^{(\theta(a(R))-\tilde{\theta})a(R) - \theta(a(R))H^{\uparrow}_{1+\theta(a(R))}(X|Y) + \tilde{\theta}H_{1+\tilde{\theta},1+\theta(a(R))}(X|Y)}\right)\Bigg]/s, \tag{127}$$

where $\theta(a) = \theta^{\uparrow}(a)$ and $a(R) = a^{\uparrow}(R)$ are the inverse functions defined in (35) and (36).

Remark 5. *Here, we discuss the possibility for extension to the continuous case. As explained in Remark 1, we can define the information quantities for the case when $\mathcal{Y}$ is continuous, but $\mathcal{X}$ is a discrete finite set. The discussions in this subsection still hold even in this continuous case. In particular, in the n-i.i.d. extension case with this continuous setting, Lemma 14 and Corollary 1 hold when the information measures are replaced by n times the single-shot information measures.*

3.3. Finite-Length Bounds for Markov Source

In this subsection, we derive several finite-length bounds for the Markov source with a computable form. Unfortunately, it is not easy to evaluate how tight those bounds are only with their formula. Their tightness will be discussed by considering the asymptotic limit in the remaining subsections of this section. Since we assume the irreducibility for the transition matrix describing the Markov chain, the following bound holds with any initial distribution.

To derive a lower bound on $-\log \bar{\mathrm{P}}_{\mathrm{s}}(M_n)$ in terms of the Rényi entropy of the transition matrix, we substitute the formula for the Rényi entropy given in Lemma 10 into Lemma 15. Then, we can derive the following achievability bound.

Theorem 6 (Direct, Ass. 1)**.** *Suppose that transition matrix W satisfies Assumption 1. Let $R := \frac{1}{n}\log M_n$. Then, for every $n \ge 1$, we have:*

$$-\log \bar{\mathrm{P}}_{\mathrm{s}}^{(n)}(M_n) \ge \sup_{-1\le\theta\le 0}\left[-\theta n R + (n-1)\theta H^{\downarrow,W}_{1+\theta}(X|Y) + \underline{\delta}(\theta)\right], \tag{128}$$

where $\underline{\delta}(\theta)$ *is given by* (91).

For the source coding without side-information, from Lemma 16 and a special case of Lemma 10, we have the following achievability bound.

Theorem 7 (Direct, no-side-information). *Let* $R := \frac{1}{n}\log M_n$. *Then, for every* $n \geq 1$, *we have:*

$$-\log \mathrm{P}_{\mathrm{e}}^{(n)}(M_n) \geq \sup_{-1<\theta\leq 0} \frac{-n\theta R + (n-1)\theta H_{1+\theta}^{W}(X) + \underline{\delta}(\theta)}{1+\theta}. \tag{129}$$

To derive an upper bound on $-\log \mathrm{P}_{\mathrm{s}}(M_n)$ in terms of the Rényi entropy of transition matrix, we substitute the formula for the Rényi entropy given in Lemma 10 for Theorem 5. Then, we have the following converse bound.

Theorem 8 (Converse, Ass. 1). *Suppose that transition matrix* W *satisfies Assumption 1. Let* $R := \frac{1}{n}\log M_n$. *For any* $H^W(X|Y) < R < H_0^{\downarrow,W}(X|Y)$, *we have:*

$$\begin{aligned} &-\log \mathrm{P}_{\mathrm{s}}^{(n)}(M_n) \qquad (130)\\ &\leq \inf_{\substack{s>0\\ -1<\tilde{\theta}<\theta(a(R))}} \Bigg[(n-1)(1+s)\tilde{\theta}\left\{H_{1+\tilde{\theta}}^{\downarrow,W}(X|Y) - H_{1+(1+s)\tilde{\theta}}^{\downarrow,W}(X|Y)\right\} + \delta_1 \\ &\quad -(1+s)\log\left(1 - 2e^{(n-1)[(\theta^W(a^W(R))-\tilde{\theta})a^W(R) - \theta^W(a^W(R))H_{1+\theta^W(a^W(R))}^{\downarrow,W}(X|Y) + \tilde{\theta}H_{1+\tilde{\theta}}^{\downarrow,W}(X|Y)] + \delta_2}\right)\Bigg] \qquad (131)\end{aligned}$$

where $\theta(a) = \theta^{\downarrow}(a)$ *and* $a(R) = a^{\downarrow}(R)$ *are the inverse functions defined by* (67) *and* (70), *respectively,*

$$\delta_1 := (1+s)\overline{\delta}(\tilde{\theta}) - \underline{\delta}((1+s)\tilde{\theta}), \tag{132}$$

$$\delta_2 := \frac{(\theta^W(a^W(R)) - \tilde{\theta})R - (1+\tilde{\theta})\underline{\delta}(\theta^W(a^W(R))) + (1+\theta^W(a^W(R)))\overline{\delta}(\tilde{\theta})}{1+\theta^W(a^W(R))}, \tag{133}$$

and $\overline{\delta}(\cdot)$ *and* $\underline{\delta}(\cdot)$ *are given by* (90) *and* (91), *respectively.*

Proof. We first use (124) of Theorem 5 for $Q_{Y^n} = P_{Y^n}$ and Lemma 10. Then, we restrict the range of $\tilde{\theta}$ as $-1 < \tilde{\theta} < \theta^W(a^W(R))$ and set $\vartheta = \frac{\theta^W(a^W(R)) - \tilde{\theta}}{1+\tilde{\theta}}$. Then, we have the assertion of the theorem. □

Next, we derive tighter bounds under Assumption 2. To derive a lower bound on $-\log \bar{\mathrm{P}}_{\mathrm{s}}(M_n)$ in terms of the Rényi entropy of the transition matrix, we substitute the formula for the Rényi entropy in Lemma 11 for Lemma 14. Then, we have the following achievability bound.

Theorem 9 (Direct, Ass. 2). *Suppose that transition matrix* W *satisfies Assumption 2. Let* $R := \frac{1}{n}\log M_n$. *Then, we have:*

$$-\log \bar{\mathrm{P}}_{\mathrm{s}}^{(n)}(M_n) \geq \sup_{-\frac{1}{2}\leq\theta\leq 0} \frac{-\theta n R + (n-1)\theta H_{1+\theta}^{\uparrow,W}(X|Y)}{1+\theta} + \underline{\xi}(\theta), \tag{134}$$

where $\underline{\xi}(\theta)$ *is given by* (98).

Finally, to derive an upper bound on $-\log \mathrm{P}_{\mathrm{s}}(M_n)$ in terms of the Rényi entropy for the transition matrix, we substitute the formula for the Rényi entropy in Lemma 12 for Theorem 5 for $Q_{Y^n} = P_{Y^n}^{(1+\theta^W(a^W(R)))}$. Then, we can derive the following converse bound.

Theorem 10 (Converse, Ass. 2)**.** *Suppose that transition matrix W satisfies Assumption 2. Let $R := \frac{1}{n}\log M_n$. For any $H^W(X|Y) < R < H_0^{\uparrow,W}(X|Y)$, we have:*

$$-\log \mathrm{P}_{\mathrm{s}}^{(n)}(M_n) \tag{135}$$

$$\le \inf_{\substack{s>0 \\ -1<\tilde{\theta}<\theta^W(a^W(R))}} \left[(n-1)(1+s)\tilde{\theta}\left\{ H^W_{1+\tilde{\theta},1+\theta^W(a^W(R))}(X|Y) - H^W_{1+(1+s)\tilde{\theta},1+\theta^W(a^W(R))}(X|Y) \right\} + \delta_1 \right.$$
$$\left. - (1+s)\log\left(1 - 2e^{(n-1)[(\theta^W(a^W(R))-\tilde{\theta})a^W(R) - \theta^W(a^W(R)) H^{\uparrow,W}_{1+\theta^W(a^W(R))}(X|Y) + \tilde{\theta} H^W_{1+\tilde{\theta},1+\theta^W(a^W(R))}(X|Y)] + \delta_2}\right) \right] / s, \tag{136}$$

where $\theta^W(a) = \theta^{\uparrow,W}(a)$ and $a^W(R) = a^{\uparrow,W}(R)$ are the inverse functions defined by (71) and (73), respectively,

$$\delta_1 := (1+s)\overline{\zeta}(\tilde{\theta}, \theta^W(a^W(R))) - \underline{\zeta}((1+s)\tilde{\theta}, \theta^W(a^W(R))), \tag{137}$$

$$\delta_2 := \frac{(\theta^W(a^W(R)) - \tilde{\theta})R - (1+\tilde{\theta})\underline{\zeta}(\theta^W(a^W(R)), \theta^W(a^W(R))) + (1+\theta^W(a^W(R)))\overline{\zeta}(\tilde{\theta}, \theta^W(a^W(R)))}{1+\theta^W(a^W(R))}, \tag{138}$$

and $\overline{\zeta}(\cdot,\cdot)$ and $\underline{\zeta}(\cdot,\cdot)$ are given by (102)–(105).

Proof. We first use (124) of Theorem 5 for $Q_{Y^n} = P_{Y^n}^{(1+\theta^W(a^W(R)))}$ and Lemma 12. Then, we restrict the range of $\tilde{\theta}$ as $-1 < \tilde{\theta} < \theta^W(a^W(R))$ and set $\vartheta = \frac{\theta^W(a^W(R)) - \tilde{\theta}}{1+\tilde{\theta}}$. Then, we have the assertion of the theorem. □

3.4. Second-Order

By applying the central limit theorem to Lemma 13 (cf. [67] (Theorem 27.4, Example 27.6)) and Lemma 18 for $Q_Y = P_Y$ and by using Theorem 2, we have the following.

Theorem 11. *Suppose that transition matrix W on $\mathcal{X} \times \mathcal{Y}$ satisfies Assumption 1. For arbitrary $\varepsilon \in (0,1)$, we have:*

$$\log M(n,\varepsilon) = \log \bar{M}(n,\varepsilon) + o(\sqrt{n}) = nH^W(X|Y) + \sqrt{\mathrm{V}^W(X|Y)}\sqrt{n}\Phi(1-\varepsilon) + o(\sqrt{n}). \tag{139}$$

Proof. The central limit theorem for the Markov process cf. [67] (Theorem 27.4, Example 27.6) guarantees that the random variable $(-\log P_{X^n|Y^n}(X^n|Y^n) - nH^W(X|Y))/\sqrt{n}$ asymptotically obeys the normal distribution with average zero and variance $\mathrm{V}^W(X|Y)$, where we use Theorem 2 to show that the limit of the variance is given by $\mathrm{V}^W(X|Y)$. Let $R = \sqrt{\mathrm{V}^W(X|Y)}\Phi^{-1}(1-\varepsilon)$. Substituting $M = e^{nH^W(X|Y) + \sqrt{n}R}$ and $\gamma = nH^W(X|Y) + \sqrt{n}R - n^{\frac{1}{4}}$ in Lemma 13, we have:

$$\lim_{n\to\infty} \bar{\mathrm{P}}_{\mathrm{s}}^{(n)}\left(e^{nH^W(X|Y)+\sqrt{n}R}\right) \le \varepsilon. \tag{140}$$

On the other hand, substituting $M = e^{nH^W(X|Y)+\sqrt{n}R}$ and $\gamma = nH^W(X|Y) + \sqrt{n}R + n^{\frac{1}{4}}$ in Lemma 18 for $Q_Y = P_Y$, we have:

$$\lim_{n\to\infty} \mathrm{P}_{\mathrm{s}}^{(n)}\left(e^{nH^W(X|Y)+\sqrt{n}R}\right) \ge \varepsilon. \tag{141}$$

Combining (140) and (141), we have the statement of the theorem. □

From the above theorem, the (first-order) compression limit of source coding with side-information for a Markov source under Assumption 1 is given by (although the compression limit of source coding

with side-information for a Markov chain is known more generally [68], we need Assumption 1 to get a single-letter characterization):

$$\lim_{n\to\infty}\frac{1}{n}\log M(n,\varepsilon) = \lim_{n\to\infty}\frac{1}{n}\log \bar{M}(n,\varepsilon) \tag{142}$$
$$= H^W(X|Y) \tag{143}$$

for any $\varepsilon \in (0,1)$. In the next subsections, we consider the asymptotic behavior of the error probability when the rate is larger than the compression limit $H^W(X|Y)$ in the moderate deviation regime and the large deviation regime, respectively.

3.5. Moderate Deviation

From Theorems 6 and 8, we have the following.

Theorem 12. *Suppose that transition matrix W satisfies Assumption 1. For arbitrary $t \in (0,1/2)$ and $\delta > 0$, we have:*

$$\lim_{n\to\infty} -\frac{1}{n^{1-2t}}\log \mathrm{P}_{\mathrm{s}}^{(n)}\left(e^{nH^W(X|Y)+n^{1-t}\delta}\right) = \lim_{n\to\infty} -\frac{1}{n^{1-2t}}\log \bar{\mathrm{P}}_{\mathrm{s}}^{(n)}\left(e^{nH^W(X|Y)+n^{1-t}\delta}\right) \tag{144}$$
$$= \frac{\delta^2}{2\mathrm{V}^W(X|Y)}. \tag{145}$$

Proof. We apply Theorems 6 and 8 to the case with $R = H^W(X|Y) + n^{-t}\delta$, i.e., $\theta(a(R)) = -n^{-1}\frac{\delta}{\mathrm{V}^W(X|Y)} + o(n^{-t})$. For the achievability part, from (88) and Theorem 6, we have:

$$-\log \mathrm{P}_{\mathrm{s}}^{(n)}(M_n) \ge \sup_{-1\le\theta\le 0}\left[-\theta n R + (n-1)\theta H^{\downarrow,W}_{1+\theta}(X|Y)\right] + \inf_{-1\le\theta\le 0}\underline{\delta}(\theta) \tag{146}$$
$$\ge n^{1-2t}\frac{\delta^2}{2\mathrm{V}^W(X|Y)} + o(n^{1-2t}). \tag{147}$$

To prove the converse part, we fix arbitrary $s > 0$ and choose $\tilde{\theta}$ to be $-n^{-t}\frac{\delta}{\mathrm{V}^W(X|Y)} + n^{-2t}$. Then, Theorem 8 implies that:

$$\limsup_{n\to\infty} -\frac{1}{n^{1-2t}}\log \mathrm{P}_{\mathrm{s}}(M_n) \le \limsup_{n\to\infty} n^{2t}\frac{1+s}{s}\tilde{\theta}\left\{H^{\downarrow,W}_{1+\tilde{\theta}}(X|Y) - H^{\downarrow,W}_{1+(1+s)\tilde{\theta}}(X|Y)\right\} \tag{148}$$
$$= \limsup_{n\to\infty} n^{2t}\frac{1+s}{s}s\tilde{\theta}^2\left.\frac{dH^{\downarrow,W}_{1+\theta}(X|Y)}{d\theta}\right|_{\theta=\tilde{\theta}} \tag{149}$$
$$= (1+s)\frac{\delta^2}{2\mathrm{V}^W(X|Y)}. \tag{150}$$

□

Remark 6. *In the literature [13,69], the moderate deviation results are stated for ϵ_n such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ instead of n^{-t} for $t \in (0,1/2)$. Although the former is slightly more general than the latter, we employ the latter formulation in Theorem 12 since the order of convergence is clearer. In fact, n^{-t} in Theorem 12 can be replaced by general ϵ_n without modifying the argument of the proof.*

3.6. Large Deviation

From Theorems 6 and 8, we have the following.

Theorem 13. *Suppose that transition matrix W satisfies Assumption 1. For $H^W(X|Y) < R$, we have:*

$$\liminf_{n\to\infty} -\frac{1}{n}\log \bar{\mathrm{P}}_{\mathrm{s}}^{(n)}\left(e^{nR}\right) \geq \sup_{-1\leq\theta\leq 0} [-\theta R + \theta H_{1+\theta}^{\downarrow,W}(X|Y)]. \tag{151}$$

On the other hand, for $H^W(X|Y) < R < H_0^{\downarrow,W}(X|Y)$, we have:

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_{\mathrm{s}}^{(n)}\left(e^{nR}\right) \leq -\theta(a(R))a(R) + \theta(a(R))H_{1+\theta(a(R))}^{\downarrow,W}(X|Y) \tag{152}$$

$$= \sup_{-1<\theta\leq 0} \frac{-\theta R + \theta H_{1+\theta}^{\downarrow,W}(X|Y)}{1+\theta}. \tag{153}$$

Proof. The achievability bound (151) follows from Theorem 6. The converse part (152) is proven from Theorem 8 as follows. We first fix $s > 0$ and $-1 < \tilde{\theta} < \theta(a(R))$. Then, Theorem 8 implies:

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_{\mathrm{s}}^{(n)}\left(e^{nR}\right) \leq \frac{1+s}{s}\tilde{\theta}\left\{H_{1+\tilde{\theta}}^{\downarrow,W}(X|Y) - H_{1+(1+s)\tilde{\theta}}^{\downarrow,W}(X|Y)\right\}. \tag{154}$$

By taking the limit $s \to 0$ and $\tilde{\theta} \to \theta(a(R))$, we have:

$$\frac{1+s}{s}\tilde{\theta}\left\{H_{1+\tilde{\theta}}^{\downarrow,W}(X|Y) - H_{1+(1+s)\tilde{\theta}}^{\downarrow,W}(X|Y)\right\} \tag{155}$$

$$= \frac{1}{s}\left(\tilde{\theta}H_{1+\tilde{\theta}}^{\downarrow,W}(X|Y) - (1+s)\tilde{\theta}H_{1+(1+s)\tilde{\theta}}^{\downarrow,W}(X|Y)\right) + \tilde{\theta}H_{1+\tilde{\theta}}^{\downarrow,W}(X|Y) \tag{156}$$

$$\to -\tilde{\theta}\frac{d[\theta H_{1+\theta}^{\downarrow,W}(X|Y)]}{d\theta}\bigg|_{\theta=\tilde{\theta}} + \tilde{\theta}H_{1+\tilde{\theta}}^{\downarrow,W}(X|Y) \quad (\text{as } s\to 0) \tag{157}$$

$$\to -\theta(a(R))\frac{d[\theta H_{1+\theta}^{\downarrow,W}(X|Y)]}{d\theta}\bigg|_{\theta=\theta(a(R))} + \theta(a(R))H_{1+\theta(a(R))}^{\downarrow,W}(X|Y) \quad (\text{as } \tilde{\theta}\to\theta(a(R))) \tag{158}$$

$$= -\theta(a(R))a(R) + \theta(a(R))H_{1+\theta(a(R))}^{\downarrow,W}(X|Y). \tag{159}$$

Thus, (152) is proven. The alternative expression (153) is derived via Lemma 9. □

Under Assumption 2, from Theorems 9 and 10, we have the following tighter bound.

Theorem 14. *Suppose that transition matrix W satisfies Assumption 2. For $H^W(X|Y) < R$, we have:*

$$\liminf_{n\to\infty} -\frac{1}{n}\log \bar{\mathrm{P}}_{\mathrm{s}}^{(n)}\left(e^{nR}\right) \geq \sup_{-\frac{1}{2}\leq\theta\leq 0} \frac{-\theta R + \theta H_{1+\theta}^{\uparrow,W}(X|Y)}{1+\theta}. \tag{160}$$

On the other hand, for $H^W(X|Y) < R < H_0^{\uparrow,W}(X|Y)$, we have:

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_{\mathrm{s}}^{(n)}\left(e^{nR}\right) \leq -\theta(a(R))a(R) + \theta(a(R))H_{1+\theta(a(R))}^{\uparrow,W}(X|Y) \tag{161}$$

$$= \sup_{-1<\theta\leq 0} \frac{-\theta R + \theta H_{1+\theta}^{\uparrow,W}(X|Y)}{1+\theta}. \tag{162}$$

Proof. The achievability bound (160) follows from Theorem 9. The converse part (161) is proven from Theorem 10 as follows. We first fix $s > 0$ and $-1 < \tilde{\theta} < \theta(a(R))$. Then, Theorem 10 implies:

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_{\mathrm{s}}^{(n)}\left(e^{nR}\right) \leq \frac{1+s}{s}\tilde{\theta}\left\{H_{1+\tilde{\theta},1+\theta(a(R))}^{W}(X|Y) - H_{1+(1+s)\tilde{\theta},1+\theta(a(R))}^{W}(X|Y)\right\}. \tag{163}$$

By taking the limit $s \to 0$ and $\tilde{\theta} \to \theta(a(R))$, we have:

$$\frac{1+s}{s}\tilde{\theta}\left\{H^W_{1+\tilde{\theta},1+\theta(a(R))}(X|Y) - H^W_{1+(1+s)\tilde{\theta},1+\theta(a(R))}(X|Y)\right\} \tag{164}$$

$$= \frac{1}{s}\left(\tilde{\theta}H^W_{1+\tilde{\theta},1+\theta(a(R))}(X|Y) - (1+s)\tilde{\theta}H^W_{1+(1+s)\tilde{\theta},1+\theta(a(R))}(X|Y)\right) + \tilde{\theta}H^W_{1+\tilde{\theta},1+\theta(a(R))}(X|Y) \tag{165}$$

$$\to -\tilde{\theta}\frac{d[\theta H^W_{1+\theta,1+\theta(a(R))}(X|Y)]}{d\theta}\bigg|_{\theta=\tilde{\theta}} + \tilde{\theta}H^W_{1+\tilde{\theta},1+\theta(a(R))}(X|Y) \quad (\text{as } s \to 0) \tag{166}$$

$$\to -\theta(a(R))\frac{d[\theta H^W_{1+\theta,1+\theta(a(R))}(X|Y)]}{d\theta}\bigg|_{\theta=\theta(a(R))} + \theta(a(R))H^{\uparrow,W}_{1+\theta(a(R))}(X|Y) \quad (\text{as } \tilde{\theta} \to \theta(a(R))) \tag{167}$$

$$= -\theta(a(R))a(R) + \theta(a(R))H^{\uparrow,W}_{1+\theta(a(R))}(X|Y). \tag{168}$$

Thus, (161) is proven. The alternative expression (162) is derived via Lemma 9. □

Remark 7. *For $R \le R_{\mathrm{cr}}$, where (cf. (72) for the definition of $R(a)$):*

$$R_{\mathrm{cr}} := R\left(\frac{d[\theta H^{\uparrow,W}_{1+\theta}(X|Y)]}{d\theta}\bigg|_{\theta=-\frac{1}{2}}\right) \tag{169}$$

is the critical rate, the left-hand side of (76) in Lemma 9 is attained by parameters in the range $-1/2 \le \theta \le 0$. Thus, the lower bound in (160) is rewritten as:

$$\sup_{-\frac{1}{2}\le\theta\le 0}\frac{-\theta R + \theta H^{\uparrow,W}_{1+\theta}(X|Y)}{1+\theta} = -\theta(a(R))a(R) + \theta(a(R))H^{\uparrow,W}_{1+\theta(a(R))}(X|Y). \tag{170}$$

Thus, the lower bound and the upper bounds coincide up to the critical rate.

Remark 8. *For the source coding without side-information, by taking the limit of Theorem 7, we have:*

$$\liminf_{n\to\infty} -\frac{1}{n}\log \bar{\mathrm{P}}^{(n)}_{\mathrm{s}}\left(e^{nR}\right) \ge \sup_{-1\le\theta\le 0}\frac{-\theta R + \theta H^W_{1+\theta}(X)}{1+\theta}. \tag{171}$$

On the other hand, as a special case of (152) without side-information, we have:

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathrm{P}^{(n)}_{\mathrm{s}}\left(e^{nR}\right) \le \sup_{-1<\theta\le 0}\frac{-\theta R + \theta H^W_{1+\theta}(X)}{1+\theta} \tag{172}$$

for $H^W(X) < R < H^W_0(X)$. Thus, we can recover the results in [40,41] by our approach.

3.7. Numerical Example

In this section, to demonstrate the advantage of our finite-length bound, we numerically evaluate the achievability bound in Theorem 7 and a special case of the converse bound in Theorem 8 for the source coding without side-information. Thanks to the aspect (A2), our numerical calculation shows that our upper finite-length bounds are very close to our lower finite-length bounds when the size n is sufficiently large. Thanks to the aspect (A1), we could calculate both bounds with the huge size $n = 1 \times 10^5$ because the calculation complexity behaves as $O(1)$.

We consider a binary transition matrix W given by Figure 2, i.e.,

$$W = \begin{bmatrix} 1-p & q \\ p & 1-q \end{bmatrix}. \tag{173}$$

In this case, the stationary distribution is:

$$\tilde{P}(0) = \frac{q}{p+q}, \tag{174}$$

$$\tilde{P}(1) = \frac{p}{p+q}. \tag{175}$$

The entropy is:

$$H^W(X) = \frac{q}{p+q}h(p) + \frac{p}{p+q}h(q), \tag{176}$$

where $h(\cdot)$ is the binary entropy function. The tilted transition matrix is:

$$W_\theta = \begin{bmatrix} (1-p)^{1+\theta} & q^{1+\theta} \\ p^{1+\theta} & (1-q)^{1+\theta} \end{bmatrix}. \tag{177}$$

The Perron–Frobenius eigenvalue is:

$$\lambda_\theta = \frac{(1-p)^{1+\theta} + (1-q)^{1+\theta} + \sqrt{\{(1-p)^{1+\theta} - (1-q)^{1+\theta}\}^2 + 4p^{1+\theta}q^{1+\theta}}}{2} \tag{178}$$

and its normalized eigenvector is:

$$\tilde{P}_\theta(0) = \frac{q^{1+\theta}}{\lambda_\theta - (1-p)^{1+\theta} + q^{1+\theta}}, \tag{179}$$

$$\tilde{P}_\theta(1) = \frac{\lambda_\theta - (1-p)^{1+\theta}}{\lambda_\theta - (1-p)^{1+\theta} + q^{1+\theta}}. \tag{180}$$

The normalized eigenvector of W_ρ^T is also given by:

$$\hat{P}_\theta(0) = \frac{p^{1+\theta}}{\lambda_\theta - (1-p)^{1+\theta} + p^{1+\theta}}, \tag{181}$$

$$\hat{P}_\theta(1) = \frac{\lambda_\theta - (1-p)^{1+\theta}}{\lambda_\theta - (1-p)^{1+\theta} + p^{1+\theta}}. \tag{182}$$

From these calculations, we can evaluate the bounds in Theorems 7 and 8. For $p = 0.1$, $q = 0.2$, the bounds are plotted in Figure 3 for fixed error probability $\varepsilon = 10^{-3}$. Although there is a gap between the achievability bound and the converse bound for rather small n, the gap is less than approximately 5% of the entropy rate for n larger than 10,000. We also plot the bounds in Figure 4 for fixed block length $n = 10{,}000$ and varying ε. The gap between the achievability bound and the converse bound remains approximately 5% of the entropy rate even for ε as small as 10^{-10}.

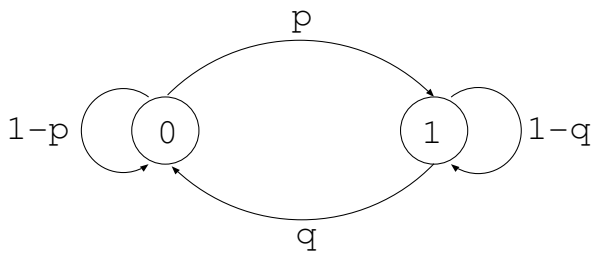

Figure 2. The description of the transition matrix in (173).

The gap between the achievability bound and the converse bound in Figure 3 is rather large compared to a similar numerical experiment conducted in [1]. One reason for the gap is that our

bounds are exponential-type bounds. For instance, when the source is i.i.d., the achievability bound essentially reduces to the so-called Gallager bound [64]. However, an advantage of our bounds is that the computational complexity does not depend on the blocklength. The computational complexities of the bounds plotted in [1] depend the blocklength, and numerical computation of those bounds for Markov sources seems to be difficult.

When $p = q$, an alternative approach to derive tighter bounds is to consider encoding of the Markov transition, i.e., $\mathbf{1}[X_i = X_{i+1}]$, instead of the source itself (cf. [45] (Example 4)). Then, the analysis can be reduced to i.i.d. case. However, such an approach is possible only when $p = q$.

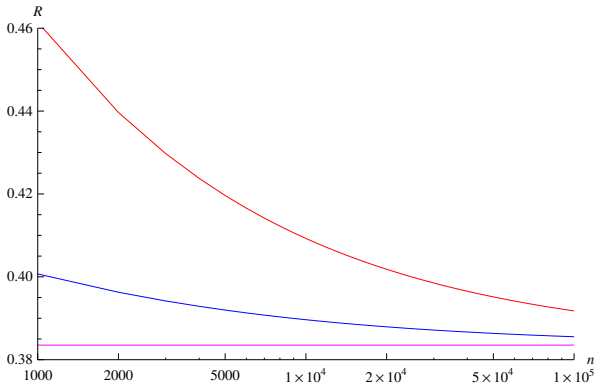

Figure 3. A comparison of the bounds for $p = 0.1$, $q = 0.2$, and $\varepsilon = 10^{-3}$. The horizontal axis is the block length n, and the vertical axis is the rate R (nats). The upper red curve is the achievability bound in Theorem 7. The middle blue curve is the converse bound in Theorem 8. The lower purple line is the first-order asymptotics given by the entropy $H^W(X)$.

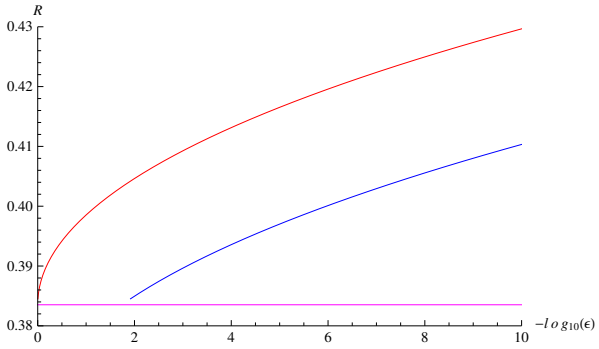

Figure 4. A comparison of the bounds for $p = 0.1$, $q = 0.2$, and $n = 10{,}000$. The horizontal axis is $-\log_{10}(\varepsilon)$, and the vertical axis is the rate R (nats). The upper red curve is the achievability bound in Theorem 7. The middle blue curve is the converse bound in Theorem 8. The lower purple line is the first-order asymptotics given by the entropy $H^W(X)$.

3.8. Summary of the Results

The obtained results in this section are summarized in Table 2. The check marks ✓ indicate that the tight asymptotic bounds (large deviation, moderate deviation, and second-order) can be obtained from those bounds. The marks $\checkmark^*$ indicate that the large deviation bound can be derived up to the critical rate. The computational complexity "Tail" indicates that the computational complexities of those bounds depend on the computational complexities of tail probabilities. It should be noted that Theorem 8 is derived from a special case ($Q_Y = P_Y$) of Theorem 5. The asymptotically optimal choice is $Q_Y = P_Y^{(1+\theta)}$, which corresponds to Corollary 1. Under Assumption 1, we can derive the bound of the Markov case only for that special choice of Q_Y, while under Assumption 2, we can derive the bound of the Markov case for the optimal choice of Q_Y.

Table 2. Summary of the bounds for source coding with full side-information. No-side means the case with no side-information.

Ach./Conv.	Markov	Single-Shot	$P_s/\bar{P}_s$	Complexity	Large Deviation	Moderate Deviation	Second Order
Achievability	Theorem 6 (Ass. 1)	Lemma 15	$\bar{P}_s$	$O(1)$		✓	
	Theorem 9 (Ass. 2)	Lemma 14	$\bar{P}_s$	$O(1)$	✓*	✓	
	Theorem 7 (No-side)	Lemma 16	$\bar{P}_s$	$O(1)$	✓*	✓	
	Lemma 13		$\bar{P}_s$	Tail		✓	✓
Converse	Theorem 8 (Ass. 1)	(Theorem 5)	P_s	$O(1)$		✓	
	Theorem 10 (Ass. 2)	Corollary 1	P_s	$O(1)$	✓*	✓	
	Theorem 8 (No-side)	(Theorem 5)	P_s	$O(1)$	✓*	✓	
	Lemma 18		P_s	Tail		✓	✓

4. Channel Coding

In this section, we investigate the channel coding with a conditional additive channel. The first part of this section discusses the general properties of the channel coding with a conditional additive channel. The second part of this section discusses the properties of the channel coding when the conditional additive noise of the channel is Markov. The first part starts with showing the problem setting in Section 4.1 by introducing a conditional additive channel. Section 4.2 gives a canonical method to convert a regular channel to a conditional additive channel. Section 4.3 gives a method to convert a BPSK-AWGN channel to a conditional additive channel. Then, we show some single-shot achievability bounds in Section 4.4 and single-shot converse bounds in Section 4.5.

As the second part, we derive finite-length bounds for the Markov noise channel in Section 4.6. Then, we derive the second-order rate in Section 4.7. In Sections 4.8 and 4.9, we show the asymptotic characterization for the large deviation regime and the moderate deviation regime by using those finite-length bounds.

4.1. Formulation for the Conditional Additive Channel

4.1.1. Single-Shot Case

We first present the problem formulation in the single-shot setting. For a channel $P_{B|A}(b|a)$ with input alphabet $\mathcal{A}$ and output alphabet $\mathcal{B}$, a channel code $\Psi = (\mathsf{e}, \mathsf{d})$ consists of one encoder $\mathsf{e} : \{1, \ldots, M\} \to \mathcal{A}$ and one decoder $\mathsf{d} : \mathcal{B} \to \{1, \ldots, M\}$. The average decoding error probability is defined by:

$$P_c[\Psi] := \sum_{m=1}^{M} \frac{1}{M} P_{B|A}(\{b : \mathsf{d}(b) \neq m\}|\mathsf{e}(m)). \tag{183}$$

For notational convenience, we introduce the error probability under the condition that the message size is M:

$$P_c(M) := \inf_{\Psi} P_c[\Psi]. \tag{184}$$

Assume that the input alphabet $\mathcal{A}$ is the same set as the output alphabet $\mathcal{B}$ and they equal an additive group $\mathcal{X}$. When the transition matrix $P_{B|A}(b|a)$ is given as $P_X(b-a)$ by using a distribution P_X on $\mathcal{X}$, the channel is called additive.

To extend the concept of the additive channel, we consider the case when the input alphabet $\mathcal{A}$ is an additive group $\mathcal{X}$ and the output alphabet $\mathcal{B}$ is the product set $\mathcal{X} \times \mathcal{Y}$. When the transition matrix $P_{B|A}(x,y|a)$ is given as $P_{XY}(x-a,y)$ by using a distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$, the channel is called conditional additive. In this paper, we are exclusively interested in the conditional additive channel. As explained in Section 4.2, a channel is a conditional additive channel if and only if it is a regular channel in the sense of [31]. When we need to express the underlying distribution of the noise explicitly, we denote the average decoding error probability by $\mathrm{P_c}[\Psi|P_{XY}]$.

4.1.2. n-Fold Extension

When we consider n-fold extension, the channel code is denoted with subscript n such as $\Psi_n = (\mathsf{e}_n, \mathsf{d}_n)$. The error probabilities given in (183) and (184) are written with the superscript (n) as $\mathrm{P_c}^{(n)}[\Psi_n]$ and $\mathrm{P_c}^{(n)}(M_n)$, respectively. Instead of evaluating the error probability $\mathrm{P_c}^{(n)}(M_n)$ for given M_n, we are also interested in evaluating:

$$M(n,\varepsilon) := \sup\left\{M_n : \mathrm{P_c}^{(n)}(M_n) \le \varepsilon\right\} \tag{185}$$

for given $0 \le \varepsilon \le 1$.

When the channel is given as a conditional distribution, the channel is given by:

$$P_{B^n|A^n}(x^n,y^n|a^n) = P_{X^nY^n}(x^n - a^n, y^n), \tag{186}$$

where $P_{X^nY^n}$ is a noise distribution on $\mathcal{X}^n \times \mathcal{Y}^n$.

For the code construction, we investigate the linear code. For an (n,k) linear code $\mathcal{C}_n \subset \mathcal{A}^n$, there exists a parity check matrix $f_n : \mathcal{A}^n \to \mathcal{A}^{n-k}$ such that the kernel of f_n is $\mathcal{C}_n$. That is, given a parity check matrix $f_n : \mathcal{A}^n \to \mathcal{A}^{n-k}$, we define the encoder $I_{\mathrm{Ker}(f_n)} : \mathcal{C}_n \to \mathcal{A}^n$ as the imbedding of the kernel $\mathrm{Ker}(f_n)$. Then, using the decoder $\mathsf{d}_{f_n} := \operatorname{argmin}_{\mathsf{d}} \mathrm{P_c}[(I_{\mathrm{Ker}(f_n)}, \mathsf{d})]$, we define $\Psi(f_n) = (I_{\mathrm{Ker}(f_n)}, \mathsf{d}_{f_n})$.

Here, we employ a randomized choice of a parity check matrix. In particular, instead of a two-universal hash function, we focus on linear two-universal hash functions, because the linearity is required in the above relation with source coding. Therefore, denoting the set of linear two-universal hash functions from $\mathcal{A}^n$ to $\mathcal{A}^{n-k}$ by $\mathcal{F}_l$, we introduce the quantity:

$$\bar{\mathrm{P}}_{\mathrm{c}}(n,k) := \sup_{F_n \in \mathcal{F}_l} \mathbb{E}_{F_n}\left[\mathrm{P_c}^{(n)}[\Psi(F_n)]\right]. \tag{187}$$

Taking the infimum over all linear codes associated with F_n (cf. (113)), we obviously have:

$$\mathrm{P_c}^{(n)}(|\mathcal{A}|^k) \le \bar{\mathrm{P}}_{\mathrm{c}}(n,k). \tag{188}$$

When we consider the error probability for conditionally additive channels, we use notation $\bar{\mathrm{P}}_{\mathrm{c}}(n,k|P_{XY})$ so that the underlying distribution of the noise is explicit. We are also interested in characterizing:

$$k(n,\varepsilon) := \sup\{k : \bar{\mathrm{P}}_{\mathrm{c}}(n,k) \le \varepsilon\} \tag{189}$$

for given $0 \le \varepsilon \le 1$.

4.2. Conversion from the Regular Channel to the Conditional Additive Channel

The aim of this subsection is to show the following theorem by presenting the conversion rule between these two types of channels. Then, we see that a binary erasure symmetric channel is an example of a regular channel.

Theorem 15. *A channel is a regular channel in the sense of [31] if and only if it can be written as a conditional additive channel.*

To show the conversion from a conditional additive channel to a regular channel, we assume that the input alphabet $\mathcal{A}$ has an additive group structure. Let $P_{\tilde{X}}$ be a distribution on the output alphabet $\mathcal{B}$. Let π_a be a representation of the group $\mathcal{A}$ on $\mathcal{B}$, and let $G = \{\pi_a : a \in \mathcal{A}\}$. A regular channel [31] is defined by:

$$P_{B|A}(b|a) = P_{\tilde{X}}(\pi_a(b)). \tag{190}$$

The group action induces orbit:

$$\mathrm{Orb}(b) := \{\pi_a(b) : a \in \mathcal{A}\}. \tag{191}$$

The set of all orbits constitutes a disjoint partition of $\mathcal{B}$. A set of the orbits is denoted by $\vec{\mathcal{B}}$, and let $\mathrm{Orb} : \mathcal{B} \to \vec{\mathcal{B}}$ be the map to the representatives.

Example 4 (Binary erasure symmetric channel)**.** *Let* $\mathcal{A} = \{0, 1\}$, $\mathcal{B} = \{0, 1, ?\}$, *and:*

$$P_{\tilde{X}}(b) = \begin{cases} 1 - p - p' & \text{if } b = 0 \\ p & \text{if } b = 1 \\ p' & \text{if } b = ? \end{cases}. \tag{192}$$

Then, let:

$$\pi_0 = \begin{bmatrix} 0 & 1 & ? \\ 0 & 1 & ? \end{bmatrix}, \quad \pi_1 = \begin{bmatrix} 0 & 1 & ? \\ 1 & 0 & ? \end{bmatrix}. \tag{193}$$

The channel defined in this way is a regular channel (see Figure 5). In this case, there are two orbits: $\{0, 1\}$ *and* $\{?\}$.

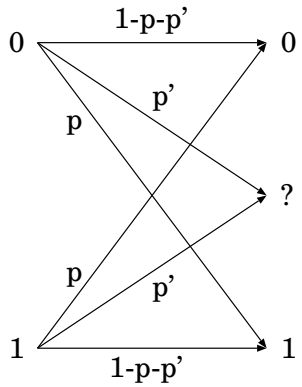

Figure 5. The binary erasure symmetric channel.

Let $\mathcal{B} = \mathcal{X} \times \mathcal{Y}$ and $P_{\tilde{X}} = P_{XY}$ for some joint distribution on $\mathcal{X} \times \mathcal{Y}$. Now, we consider a conditional additive channel, whose transition matrix $P_{B|A}(x, y|a)$ is given as $P_{XY}(x - a, y)$. When the

group action is given by $\pi_a(x,y) = (x-a, y)$, the above conditional additive channel is given as a regular channel. In this case, there are $|\mathcal{Y}|$ orbits, and the size of each orbit is $|\mathcal{X}|$, respectively. This fact shows that any conditional additive channel is written as a regular channel. That is, it shows the "if" part of Theorem 15.

Conversely, we present the conversion from a regular channel to a conditional additive channel. We first explain the construction for the single-shot channel. For random variable $\tilde{X} \sim P_{\tilde{X}}$, let $\mathcal{Y} = \bar{\mathcal{B}}$ and $Y = \varpi(\tilde{X})$ be the random variable describing the representatives of the orbits. For $y = \mathrm{Orb}(b)$ and each orbit $\mathrm{Orb}(b)$, we fix an element $0_y \in \mathrm{Orb}(b)$. Then, we define:

$$P_Y(y) := P_{\tilde{X}}(\mathrm{Orb}(b)), \quad P_{X,Y}(a,y) := \frac{P_{\tilde{X}}(\pi_a(0_y))}{|\{a' \in \mathcal{A} | \pi_a(0_y) = \pi_{a'}(0_y)\}|}. \tag{194}$$

Then, we obtain the virtual channel $P_{X,Y|A}$ as $P_{X,Y|A}(x,y|a) := P_{X,Y}(x-a,y)$. Using the conditional distributions $P_{X,Y|B}$ and $P_{B|X,Y}$ as:

$$P_{X,Y|B}(a,y|b) = \begin{cases} \frac{1}{|\{a' \in \mathcal{A} | \pi_a(0_y) = \pi_{a'}(0_y)\}|} & \text{when } b = \pi_a(0_y) \\ 0 & \text{otherwise.} \end{cases} \tag{195}$$

$$P_{X,Y|B}(a,y|b) = \begin{cases} 1 & \text{when } b = \pi_a(0_y) \\ 0 & \text{otherwise,} \end{cases} \tag{196}$$

we obtain the relations:

$$P_{B|A}(b|a) = \sum_{x,y} P_{B|X,Y}(b|x,y) P_{X,Y|A}(x,y|a), \quad P_{X,Y|A}(x,y|a) = \sum_b P_{X,Y|B}(x,y|b) P_{B|A}(b|a). \tag{197}$$

These two equations show that the receiver information of the virtual conditional additive channel $P_{X,Y|A}$ and the receiver information of the regular channel $P_{B|A}$ can be converted into each other. Hence, we can say that a regular channel in the sense of [31] can be written as a conditional additive channel, which shows the "only if" part of Theorem 15.

Example 5 (Binary erasure symmetric channel revisited)**.** *We convert the regular channel of Example 4 to a conditional additive channel. Let us label the orbit* $\{0,1\}$ *as* $y = 0$ *and* $\{?\}$ *as* $y = 1$*. Let* $0_0 = 0$ *and* $0_1 = ?$*.*

$$P_{X,Y}(x,0) = \begin{cases} 1-p-p' & \text{if } x = 0 \\ p & \text{if } x = 0 \end{cases} \tag{198}$$

$$P_{X,Y}(x,1) = \frac{p'}{2}. \tag{199}$$

When we consider the nth extension, a channel is given by:

$$P_{B^n|A^n}(b^n|a^n) = P_{\tilde{X}^n}(\pi_{a^n}(b^n)), \tag{200}$$

where the nth extension of the group action is defined by $\pi_{a^n}(b^n) = (\pi_{a_1}(b_1), \ldots, \pi_{a_n}(b_n))$.

Similarly, for n-fold extension, we can also construct the virtual conditional additive channel. More precisely, for $\tilde{X}^n \sim P_{\tilde{X}^n}$, we set $Y^n = \varpi(\tilde{X}^n) = (\varpi(\tilde{X}_1), \ldots, \varpi(\tilde{X}_n))$ and:

$$P_{X^n,Y^n}(x^n,y^n) := \frac{P_{\tilde{X}^n}(\pi_{a^n}(0_{y^n}))}{|\{a'^n \in \mathcal{A}^n | \pi_{a^n}(0_{y^n}) = \pi_{a'^n}(0_{y^n})\}|}. \tag{201}$$

4.3. Conversion of the BPSK-AWGN Channel into the Conditional Additive Channel

Although we only considered finite input/output sources and channels throughout the paper, in order to demonstrate the utility of the conditional additive channel framework, let us consider the additive white Gaussian noise (AWGN) channel with binary phase shift keying (BPSK) in this section. Let $\mathcal{A} = \{0, 1\}$ be the input alphabet of the channel, and let $\mathcal{B} = \mathbb{R}$ be the output alphabet of the channel. For an input $a \in \mathcal{A}$ and Gaussian noise Z with mean zero and variance σ^2, the output of the channel is given by $B = (-1)^a + Z$. Then, the conditional probability density function of this channel is given as:

$$P_{B|A}(b|a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(b-(-1)^a)^2}{\sigma^2}}. \tag{202}$$

Now, to define a conditional additive channel, we choose $\mathcal{Y} := \mathbb{R}_+$ and define the probability density function p_Y on $\mathcal{Y}$ with respect to the Lebesgue measure and the conditional distribution $P_{X|Y}(x|y)$ as:

$$p_Y(y) := \frac{1}{\sqrt{2\pi}\sigma}(e^{-\frac{(y-1)^2}{\sigma^2}} + e^{-\frac{(y+1)^2}{\sigma^2}}) \tag{203}$$

$$P_{X|Y}(0|y) := \frac{e^{-\frac{(y-1)^2}{\sigma^2}}}{e^{-\frac{(y-1)^2}{\sigma^2}} + e^{-\frac{(y+1)^2}{\sigma^2}}} \tag{204}$$

$$P_{X|Y}(1|y) := \frac{e^{-\frac{(y+1)^2}{\sigma^2}}}{e^{-\frac{(y-1)^2}{\sigma^2}} + e^{-\frac{(y+1)^2}{\sigma^2}}} \tag{205}$$

for $y \in \mathbb{R}_+$. When we define $b := (-1)^x y \in \mathbb{R}$ for $x \in \{0, 1\}$ and $y \in \mathbb{R}_+$, we have:

$$p_{XY|A}(y, x|a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(-1)^{a+x})^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{((-1)^x y-(-1)^a)^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(b-(-1)^a)^2}{\sigma^2}}. \tag{206}$$

The relations (202) and (206) show that the AWGN channel with BPSK is given as a conditional additive channel in the above sense.

By noting this observation, as explained in Remark 5, the single-shot achievability bounds in Section 3.2 are also valid for continuous Y. Furthermore, the discussions for the single-shot converse bounds in Section 4.5 hold even for continuous Y. Therefore, the bounds in Sections 4.4 and 4.5 are also applicable to the BPSK-AWGN channel.

In particular, in the n memoryless extension of the BPSK-AWGN channel, the information measures for the noise distribution are given as n times the single-shot information measures for the noise distribution. Even in this case, the upper and lower bounds in Sections 4.4 and 4.5 are also applicable by replacing the information measures by n times the single-shot information measures. Therefore, we obtain finite-length upper and lower bounds of the optimal coding length for the memoryless BPSK-AWGN channel. Furthermore, even though the additive noise is not Gaussian, when the probability density function p_Z of the additive noise Z satisfies the symmetry $p_Z(z) = p_Z(-z)$, the BPSK channel with the additive noise Z can be converted to a conditional additive channel in the same way.

4.4. Achievability Bound Derived by Source Coding with Side-Information

In this subsection, we give a code for a conditional additive channel from a code of source coding with side-information in a canonical way. In this construction, we see that the decoding error probability of the channel code equals that of the source code.

When the channel is given as the conditional additive channel with conditional additive noise distribution $P_{X^nY^n}$ as (186) and $\mathcal{X} = \mathcal{A}$ is the finite field $\mathbb{F}_q$, we can construct a linear channel code from a source code with full side-information whose encoder and decoder are f_n and d_n as follows. First, we assume linearity for the source encoder f_n. Let $\mathcal{C}_n(f_n)$ be the kernel of the linear encoder f_n of the source code. Suppose that the sender sends a codeword $c_n \in \mathcal{C}_n(f_n)$ and $(c_n + X^n, Y^n)$ is received. Then, the receiver computes the syndrome $f_n(c_n + X^n) = f_n(c_n) + f_n(X^n) = f_n(X^n)$, estimates X^n from $f_n(X^n)$ and Y^n, and subtracts the estimate from $c_n + X^n$. That is, we choose the channel decoder $\tilde{\mathsf{d}}_n$ as:

$$\tilde{\mathsf{d}}_n(x'^n, y^n) := x'^n - \mathsf{d}_n(f_n(x'^n), y^n). \tag{207}$$

We succeed in decoding in this channel coding if and only if $\mathsf{d}_n(f_n(X^n), Y^n)$ equals X^n. Thus, the error probability of this channel code coincides with that of the source code for the correlated source (X^n, Y^n). In summary, we have the following lemma, which was first pointed out in [27].

Lemma 19 ([27, (19)])**.** *Given a linear encoder f_n and a decoder d_n for a source code with side-information with distribution $P_{X^nY^n}$, let $I_{\mathrm{Ker}(f_n)}$ and $\tilde{\mathsf{d}}_n$ be the channel encoder and decoder induced by (f_n, d_n). Then, the error probability of channel coding for the conditionally additive channel with noise distribution $P_{X^nY^n}$ satisfies:*

$$\mathrm{P}_{\mathrm{c}}^{(n)}[(I_{\mathrm{Ker}(f_n)}, \tilde{\mathsf{d}}_n)|P_{X^nY^n}] = \mathrm{P}_{\mathrm{s}}^{(n)}[(f_n, \mathsf{d}_n)|P_{X^nY^n}]. \tag{208}$$

Furthermore, (in fact, when we additionally impose the linearity on the random function F in the definition (114) for the definition of $\bar{\mathrm{P}}_{\mathrm{s}}(M|P_{X^nY^n})$, the result in [27] implies that the equality in (209) holds) taking the infimum for F_n chosen to be a linear two-universal hash function, we also have:

$$\begin{aligned}\bar{\mathrm{P}}_{\mathrm{c}}(n,k) &= \sup_{F_n \in \mathcal{F}_l} \mathbb{E}_{F_n}\left[\mathrm{P}_{\mathrm{c}}^{(n)}[\Psi(F_n)]\right] \le \sup_{F_n \in \mathcal{F}_l} \mathbb{E}_{F_n}\left[\mathrm{P}_{\mathrm{c}}^{(n)}[(I_{\mathrm{Ker}(F_n)}, \tilde{\mathsf{d}}_n)]\right] \\ &= \sup_{F_n \in \mathcal{F}_l} \mathbb{E}_{F_n} \mathrm{P}_{\mathrm{s}}^{(n)}[(F_n, \mathsf{d}_n)] \le \sup_{F_n \in \mathcal{F}} \mathbb{E}_{F_n} \mathrm{P}_{\mathrm{s}}^{(n)}[(F_n, \mathsf{d}_n)] = \bar{\mathrm{P}}_{\mathrm{s}}^{(n)}(|\mathcal{A}^{n-k}|).\end{aligned} \tag{209}$$

By using this observation and the results in Section 3.2, we can derive the achievability bounds. By using the conversion argument in Section 4.2, we can also construct a channel code for a regular channel from a source code with full side-information. Although the following bounds are just a specialization of known bounds for conditional additive channels, we review these bounds here to clarify the correspondence between the bounds in source coding with side-information and channel coding.

From Lemma 13 and (209), we have the following.

Lemma 20 ([2])**.** *The following bound holds:*

$$\bar{\mathrm{P}}_{\mathrm{c}}(n,k) \le \inf_{\gamma \ge 0} \left[P_{X^nY^n}\left\{ \log \frac{1}{P_{X^n|Y^n}(x^n|y^n)} > \gamma \right\} + \frac{e^\gamma}{|\mathcal{A}|^{n-k}} \right]. \tag{210}$$

From Lemma 14 and (209), we have the following exponential-type bound.

Lemma 21 ([6])**.** *The following bound holds:*

$$\bar{\mathrm{P}}_{\mathrm{c}}(n,k) \le \inf_{-\frac{1}{2} \le \theta \le 0} |\mathcal{A}|^{\frac{\theta(n-k)}{1+\theta}} e^{-\frac{\theta}{1+\theta} H_{1+\theta}^{\uparrow}(X^n|Y^n)}. \tag{211}$$

From Lemma 15 and (209), we have the following slightly loose exponential bound.

Lemma 22 ([3,70]). *The following bound holds (The bound (212) was derived in the original Japanese edition of [3], but it is not written in the English edition [3]. The quantum analogue was derived in [70].):*

$$\bar{\mathrm{P}}_{\mathrm{c}}(n,k) \le \inf_{-1\le\theta\le 0} |\mathcal{A}|^{\theta(n-k)} e^{-\theta H^{\downarrow}_{1+\theta}(X^n|Y^n)}. \tag{212}$$

When X has no side-information, i.e., the virtual channel is additive, we have the following special case of Lemma 21.

Lemma 23 ([6]). *Suppose that X has no side-information. Then, the following bound holds:*

$$\bar{\mathrm{P}}_{\mathrm{c}}(n,k) \le \inf_{-\frac{1}{2}\le\theta\le 0} |\mathcal{A}|^{\frac{\theta(n-k)}{1+\theta}} e^{-\frac{\theta}{1+\theta} H_{1+\theta}(X^n)}. \tag{213}$$

4.5. Converse Bound

In this subsection, we show some converse bounds. The following is the information spectrum-type converse shown in [4].

Lemma 24 ([4], Lemma 4). *For any code $\Psi_n = (\mathsf{e}_n, \mathsf{d}_n)$ and any output distribution $Q_{B^n} \in \mathcal{P}(\mathcal{B}^n)$, we have:*

$$\mathrm{P}_{\mathrm{c}}^{(n)}[\Psi_n] \ge \sup_{\gamma\ge 0} \left[\sum_{m=1}^{M_n} \frac{1}{M_n} P_{B^n|A^n} \left\{ \log \frac{P_{B^n|A^n}(b^n|\mathsf{e}_n(m))}{Q_{B^n}(b^n)} < \gamma \right\} - \frac{e^{\gamma}}{M_n} \right]. \tag{214}$$

When a channel is a conditional additive channel, we have:

$$P_{B^n|A^n}(a^n + x^n, y^n|a^n) = P_{X^nY^n}(x^n, y^n). \tag{215}$$

By taking the output distribution Q_{B^n} as:

$$Q_{B^n}(a^n + x^n, y^n) = \frac{1}{|\mathcal{A}|^n} Q_{Y^n}(y^n) \tag{216}$$

for some $Q_{Y^n} \in \mathcal{P}(\mathcal{Y}^n)$, as a corollary of Lemma 24, we have the following bound.

Lemma 25. *When a channel is a conditional additive channel, for any distribution $Q_{Y^n} \in \mathcal{P}(\mathcal{Y}^n)$, we have:*

$$\mathrm{P}_{\mathrm{c}}^{(n)}(M_n) \ge \sup_{\gamma\ge 0} \left[P_{X^nY^n} \left\{ \log \frac{Q_{Y^n}(y^n)}{P_{X^nY^n}(x^n, y^n)} > n \log |\mathcal{A}| - \gamma \right\} - \frac{e^{\gamma}}{M_n} \right]. \tag{217}$$

Proof. By noting (215) and (216), the first term of the right-hand side of (214) can be rewritten as:

$$\sum_{m=1}^{M_n} \frac{1}{M_n} P_{B^n|A^n} \left\{ \log \frac{P_{B^n|A^n}(b^n|\mathsf{e}_n(m))}{Q_{B^n}(b^n)} < \gamma \right\} \tag{218}$$

$$= \sum_{m=1}^{M_n} \frac{1}{M_n} P_{X^nY^n} \left\{ \log \frac{P_{B^n|A^n}(\mathsf{e}_n(m) + x^n, y^n|\mathsf{e}_n(m))}{Q_{B^n}(\mathsf{e}_n(m) + x^n, y^n)} \right\} \tag{219}$$

$$= P_{X^nY^n} \left\{ \log \frac{Q_{Y^n}(y^n)}{P_{X^nY^n}(x^n, y^n)} > n \log |\mathcal{A}| - \gamma \right\}, \tag{220}$$

which implies the statement of the lemma. □

A similar argument as in Theorem 5 also derives from the following converse bound.

Theorem 16. *For any $Q_{Y^n} \in \mathcal{P}(\mathcal{Y}^n)$, we have:*

$$-\log \mathrm{P}_{\mathrm{c}}^{(n)}(M_n) \tag{221}$$

$$\leq \inf_{\substack{s>0 \\ \tilde{\theta}\in\mathbb{R},\vartheta\geq 0}} \Bigg[(1+s)\tilde{\theta}\left\{H_{1+\tilde{\theta}}(P_{X^nY^n}|Q_{Y^n}) - H_{1+(1+s)\tilde{\theta}}(P_{X^nY^n}|Q_{Y^n})\right\} - (1+s)\log\left(1-2e^{-\frac{-\vartheta R+(\tilde{\theta}+\vartheta(1+\tilde{\theta}))H_{1+\tilde{\theta}+\vartheta(1+\tilde{\theta})}(P_{X^nY^n}|Q_{Y^n})-(1+\vartheta)\tilde{\theta}H_{1+\tilde{\theta}}(P_{X^nY^n}|Q_{Y^n})}{1+\vartheta}}\right)\Bigg]/s \tag{222}$$

$$\leq \inf_{\substack{s>0 \\ -1<\tilde{\theta}<\theta(a(R))}} \Bigg[(1+s)\tilde{\theta}\left\{H_{1+\tilde{\theta}}(P_{X^nY^n}|Q_{Y^n}) - H_{1+(1+s)\tilde{\theta}}(P_{X^nY^n}|Q_{Y^n})\right\} - (1+s)\log\left(1-2e^{(\theta(a(R))-\tilde{\theta})a(R)-\theta(a(R))H_{1+\theta(a(R))}(P_{X^nY^n}|Q_{Y^n})+\tilde{\theta}H_{1+\tilde{\theta}}(P_{X^nY^n}|Q_{Y^n})}\right)\Bigg]/s, \tag{223}$$

where $R = n\log|\mathcal{A}| - \log M_n$, and $\theta(a)$ and $a(R)$ are the inverse functions defined in (29) and (32), respectively.

Proof. See Appendix L. □

4.6. Finite-Length Bound for the Markov Noise Channel

From this section, we address the conditional additive channel whose conditional additive noise is subject to the Markov chain. Here, the input alphabet $\mathcal{A}^n$ equals the additive group $\mathcal{X}^n = \mathbb{F}_q^n$, and the output alphabet $\mathcal{B}^n$ is $\mathcal{X}^n \times \mathcal{Y}^n$. That is, the transition matrix describing the channel is given by using a transition matrix W on $\mathcal{X}^n \times \mathcal{Y}^n$ and an initial distribution Q as:

$$P_{B^n|A^n}(x^n + a^n, y^n|a^n) = Q(x_1, y_1)\prod_{i=2}^{n} W(x_i, y_i|x_{i-1}, y_{i-1}). \tag{224}$$

As in Section 2.2, we consider two assumptions on the transition matrix W of the noise process $(\mathbf{X}, \mathbf{Y})$, i.e., Assumptions 1 and 2. We also use the same notations as in Section 2.2.

Example 6 (Gilbert–Elliot channel with state-information available at the receiver)**.** *The Gilbert–Elliot channel [29,30] is characterized by a channel state Y^n on $\mathcal{Y}^n = \{0,1\}^n$ and an additive noise X^n on $\mathcal{X}^n = \{0,1\}^n$. The noise process (X^n, Y^n) is a Markov chain induced by the transition matrix W introduced in Example 3. For the channel input a^n, the channel output is given by $(a^n + X^n, Y^n)$ when the state-information is available at the receiver. Thus, this channel can be regarded as a conditional additive channel, and the transition matrix of the noise process satisfies Assumption 2.*

Proofs of the following bounds are almost the same as those in Section 3.3, and thus omitted. The combination of Lemmas 10 and 22 derives the following achievability bound.

Theorem 17 (Direct, Ass. 1)**.** *Suppose that the transition matrix W of the conditional additive noise satisfies Assumption 1. Let $R := \frac{n-k}{n}\log|\mathcal{A}|$. Then, we have:*

$$-\log \bar{\mathrm{P}}_{\mathrm{c}}(n,k) \quad \geq \quad \sup_{-1\leq\theta\leq 0}\left[-\theta nR + (n-1)\theta H_{1+\theta}^{\downarrow,W}(X|Y) + \underline{\delta}(\theta)\right]. \tag{225}$$

Theorem 16 for $Q_{Y^n} = P_{Y^n}$ and Lemma 10 yield the following converse bound.

Theorem 18 (Converse, Ass. 1). *Suppose that transition matrix W of the conditional additive noise satisfies Assumption 1. Let $R := \log|\mathcal{A}| - \frac{1}{n}\log M_n$. If $H^W(X|Y) < R < H_0^{\downarrow,W}(X|Y)$, then we have:*

$$-\log \mathrm{P}_{\mathrm{c}}^{(n)}(M_n) \tag{226}$$

$$\leq \inf_{\substack{s>0 \\ -1<\tilde{\theta}<\theta(a(R))}} \Bigg[(n-1)(1+s)\tilde{\theta}\left\{H_{1+\tilde{\theta}}^{\downarrow,W}(X|Y) - H_{1+(1+s)\tilde{\theta}}^{\downarrow,W}(X|Y)\right\} + \delta_1 - (1+s)\log\left(1 - 2e^{(n-1)[(\theta(a(R))-\tilde{\theta})a(R) - \theta(a(R))H_{1+\theta(a(R))}^{\downarrow,W}(X|Y) + \tilde{\theta}H_{1+\tilde{\theta}}^{\downarrow,W}(X|Y)] + \delta_2}\right)\Bigg]/s, \tag{227}$$

where $\theta(a) = \theta^{\downarrow}(a)$ and $a(R) = a^{\downarrow}(R)$ are the inverse functions defined by (67) and (70), respectively, and:

$$\delta_1 := (1+s)\overline{\delta}(\tilde{\theta}) - \underline{\delta}((1+s)\tilde{\theta}), \tag{228}$$

$$\delta_2 := \frac{(\theta(a(R)) - \tilde{\theta})R - (1+\tilde{\theta})\underline{\delta}(\theta(a(R))) + (1+\theta(a(R)))\overline{\delta}(\tilde{\theta})}{1+\theta(a(R))}. \tag{229}$$

Next, we derive tighter bounds under Assumption 2. From Lemmas 11 and 21, we have the following achievability bound.

Theorem 19 (Direct, Ass. 2). *Suppose that the transition matrix W of the conditional additive noise satisfies Assumption 2. Let $R := \frac{n-k}{n}\log|\mathcal{A}|$. Then, we have:*

$$-\log \bar{\mathrm{P}}_{\mathrm{c}}(n,k) \geq \sup_{-\frac{1}{2}\leq\theta\leq 0} \frac{-\theta nR + (n-1)\theta H_{1+\theta}^{\uparrow,W}(X|Y)}{1+\theta} + \underline{\xi}(\theta). \tag{230}$$

By using Theorem 16 for $Q_{Y^n} = P_{Y^n}^{(1+\theta(a(R)))}$ and Lemma 12, we obtain the following converse bound.

Theorem 20 (Converse, Ass. 2). *Suppose that the transition matrix W of the conditional additive noise satisfies Assumption 2. Let $R := \log|\mathcal{A}| - \frac{1}{n}\log M_n$. If $H^W(X|Y) < R < H_0^{\uparrow,W}(X|Y)$, we have:*

$$-\log \mathrm{P}_{\mathrm{c}}^{(n)}(M_n) \tag{231}$$

$$\leq \inf_{\substack{s>0 \\ -1<\tilde{\theta}<\theta(a(R))}} \Bigg[(n-1)(1+s)\tilde{\theta}\left\{H_{1+\tilde{\theta},1+\theta(a(R))}^{W}(X|Y) - H_{1+(1+s)\tilde{\theta},1+\theta(a(R))}^{W}(X|Y)\right\} + \delta_1 - (1+s)\log\left(1 - 2e^{(n-1)[(\theta(a(R))-\tilde{\theta})a(R) - \theta(a(R))H_{1+\theta(a(R))}^{\uparrow,W}(X|Y) + \tilde{\theta}H_{1+\tilde{\theta},1+\theta(a(R))}^{W}(X|Y)] + \delta_2}\right)\Bigg]/s, \tag{232}$$

where $\theta(a) = \theta^{\uparrow}(a)$ and $a(R) = a^{\uparrow}(R)$ are the inverse functions defined by (71) and (73), respectively, and:

$$\delta_1 := (1+s)\overline{\zeta}(\tilde{\theta}, \theta(a(R))) - \underline{\zeta}((1+s)\tilde{\theta}, \theta(a(R))), \tag{233}$$

$$\delta_2 := \frac{(\theta(a(R)) - \tilde{\theta})R - (1+\tilde{\theta})\underline{\zeta}(\theta(a(R)), \theta(a(R))) + (1+\theta(a(R)))\overline{\zeta}(\tilde{\theta}, \theta(a(R)))}{1+\theta(a(R))}. \tag{234}$$

Finally, when X has no side-information, i.e., the channel is additive, we obtain the following achievability bound from Lemma 23.

Theorem 21 (Direct, no-side-information). *Let $R := \frac{n-k}{n}\log|\mathcal{A}|$. Then, we have:*

$$-\log \bar{\mathrm{P}}_{\mathrm{c}}(n,k) \geq \sup_{-\frac{1}{2}\leq\theta\leq 0} \frac{-\theta nR + (n-1)\theta H_{1+\theta}^{W}(X) + \underline{\delta}(\theta)}{1+\theta}. \tag{235}$$

Remark 9. *Our treatment for the Markov conditional additive channel covers Markov regular channels because Markov regular channels can be reduced to Markov conditional additive channels as follows. Let* $\tilde{\mathbf{X}} = \{\tilde{X}^n\}_{n=1}^{\infty}$ *be a Markov chain on* $\mathcal{B}$ *whose distribution is given by:*

$$P_{\tilde{X}^n}(\tilde{x}^n) = Q(\tilde{x}_1)\prod_{i=2}^{n}\tilde{W}(\tilde{x}_i|\tilde{x}_{i-1}) \tag{236}$$

for a transition matrix $\tilde{W}$ *and an initial distribution* Q*. Let* $(\mathbf{X}, \mathbf{Y}) = \{(X^n, Y^n)\}_{n=1}^{\infty}$ *be the noise process of the conditional additive channel derived from the noise process* $\tilde{\mathbf{X}}$ *of the regular channel by the argument of Section 4.2. Since we can write:*

$$P_{X^nY^n}(x^n, y^n) = Q(\iota_{y_1}^{-1}(\vartheta_{y_1}(x_1)))\frac{1}{|\mathrm{Stb}(0_{y_1})|}\prod_{i=2}^{n}\tilde{W}(\iota_{y_i}^{-1}(\vartheta_{y_i}(x_i))|\iota_{y_{i-1}}^{-1}(\vartheta_{y_{i-1}}(x_{i-1})))\frac{1}{|\mathrm{Stb}(0_{y_i})|}, \tag{237}$$

the process $(\mathbf{X}, \mathbf{Y})$ *is also a Markov chain. Thus, the regular channel given by* $\tilde{\mathbf{X}}$ *is reduced to the conditional additive channel given by* $(\mathbf{X}, \mathbf{Y})$*.*

4.7. Second-Order

To discuss the asymptotic performance, we introduce the quantity:

$$C := \log|\mathcal{A}| - H^W(X|Y). \tag{238}$$

By applying the central limit theorem (cf. [67] (Theorem 27.4, Example 27.6)) to Lemmas 20 and 25 for $Q_{Y^n} = P_{Y^n}$, and by using Theorem 2, we have the following.

Theorem 22. *Suppose that the transition matrix* W *of the conditional additive noise satisfies Assumption 1. For arbitrary* $\varepsilon \in (0,1)$*, we have:*

$$\log M(n,\varepsilon) = k(n,\varepsilon)\log|\mathcal{A}| = Cn + \sqrt{\mathrm{V}^W(X|Y)}\Phi^{-1}(\varepsilon)\sqrt{n} + o(\sqrt{n}). \tag{239}$$

Proof. This theorem follows in the same manner as the proof of Theorem 11 by replacing Lemma 13 with Lemma 20 (achievability) and Lemma 18 with Lemma 25 (converse). □

From the above theorem, the (first-order) capacity of the conditional additive channel under Assumption 1 is given by:

$$\lim_{n\to\infty}\frac{1}{n}\log M(n,\varepsilon) = \lim_{n\to\infty}\frac{1}{n}\log\frac{k(n,\varepsilon)\log|\mathcal{A}|}{n} = C \tag{240}$$

for every $0 < \varepsilon < 1$. In the next subsections, we consider the asymptotic behavior of the error probability when the rate is smaller than the capacity in the moderate deviation regime and the large deviation regime, respectively.

4.8. Moderate Deviation

From Theorems 17 and 18, we have the following.

Theorem 23. *Suppose that the transition matrix* W *of the conditional additive noise satisfies Assumption 1. For arbitrary* $t \in (0, 1/2)$ *and* $\delta > 0$*, we have:*

$$\lim_{n\to\infty} -\frac{1}{n^{1-2t}}\log \mathrm{P}_{\mathrm{c}}^{(n)}\left(e^{nC-n^{1-t}\delta}\right) = \lim_{n\to\infty} -\frac{1}{n^{1-2t}}\log \bar{\mathrm{P}}_{\mathrm{c}}^{(n)}\left(n, \frac{nC-n^{1-t}\delta}{\log|\mathcal{A}|}\right) \tag{241}$$

$$= \frac{\delta^2}{2\mathrm{V}^W(X|Y)}. \tag{242}$$

Proof. The theorem follows in the same manner as Theorem 12 by replacing Theorem 6 with Theorem 17 (achievability) and Theorem 8 with Theorem 18 (converse). □

4.9. Large Deviation

From Theorem 17 and Theorem 18, we have the following.

Theorem 24. *Suppose that the transition matrix W of the conditional additive noise satisfies Assumption 1. For $H^W(X|Y) < R$, we have:*

$$\liminf_{n\to\infty} -\frac{1}{n}\log \bar{\mathrm{P}}_{\mathrm{c}}^{(n)}\left(n, n\left(1-\frac{R}{\log|\mathcal{A}|}\right)\right) \geq \sup_{-1\leq\theta\leq 0}\left[-\theta R + \theta H_{1+\theta}^{\downarrow,W}(X|Y)\right]. \tag{243}$$

On the other hand, for $H^W(X|Y) < R < H_0^{\downarrow,W}(X|Y)$, we have:

$$\begin{aligned}\limsup_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_{\mathrm{c}}^{(n)}\left(e^{n(\log|\mathcal{A}|-R)}\right) &\leq -\theta(a(R))a(R) + \theta(a(R))H_{1+\theta(a(R))}^{\downarrow,W}(X|Y) &(244)\\ &= \sup_{-1<\theta\leq 0}\frac{-\theta R + \theta H_{1+\theta}^{\downarrow,W}(X|Y)}{1+\theta}. &(245)\end{aligned}$$

Proof. The theorem follows in the same manner as Theorem 13 by replacing Theorem 6 with Theorem 17 (achievability) and Theorem 8 with Theorem 18 (converse). □

Under Assumption 2, from Theorems 19 and 20, we have the following tighter bound.

Theorem 25. *Suppose that the transition matrix W of the conditional additive noise satisfies Assumption 2. For $H^W(X|Y) < R$, we have:*

$$\liminf_{n\to\infty} -\frac{1}{n}\log \bar{\mathrm{P}}_{\mathrm{c}}^{(n)}\left(n, n\left(1-\frac{R}{\log|\mathcal{A}|}\right)\right) \geq \sup_{-\frac{1}{2}\leq\theta\leq 0}\frac{-\theta R + \theta H_{1+\theta}^{\uparrow,W}(X|Y)}{1+\theta}. \tag{246}$$

On the other hand, for $H^W(X|Y) < R < H_0^{\uparrow,W}(X|Y)$, we have:

$$\begin{aligned}\limsup_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_{\mathrm{c}}^{(n)}\left(e^{n(\log|\mathcal{A}|-R)}\right) &\leq -\theta(a(R))a(R) + \theta(a(R))H_{1+\theta(a(R))}^{\uparrow,W}(X|Y) &(247)\\ &= \sup_{-1<\theta\leq 0}\frac{-\theta R + \theta H_{1+\theta}^{\uparrow,W}(X|Y)}{1+\theta}. &(248)\end{aligned}$$

Proof. The theorem follows the same manner as Theorem 14 by replacing Theorem 9 with Theorem 19 and Theorem 10 with Theorem 20. □

When X has no side-information, i.e., the channel is additive, from Theorem 21 and (245), we have the following.

Theorem 26. *For $H^W(X) < R$, we have:*

$$\liminf_{n\to\infty} -\frac{1}{n}\log \bar{\mathrm{P}}_{\mathrm{c}}^{(n)}\left(n, n\left(1-\frac{R}{\log|\mathcal{A}|}\right)\right) \geq \sup_{-\frac{1}{2}\leq\theta\leq 0}\frac{-\theta R + \theta H_{1+\theta}^{W}(X)}{1+\theta}. \tag{249}$$

On the other hand, for $H^W(X) < R < H_0^W(X)$*, we have:*

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathrm{P}_{\mathrm{c}}^{(n)}\left(e^{n(\log|\mathcal{A}|-R)}\right) \le \sup_{-1<\theta\le 0} \frac{-\theta R + \theta H_{1+\theta}^W(X)}{1+\theta}. \tag{250}$$

Proof. The first claim follows by taking the limit of Theorem 21, and the second claim follows as a special case of (245) without side-information. □

4.10. Summary of the Results

The results shown in this section for the Markov conditional additive noise are summarized in Table 3. The check marks ✓ indicate that the tight asymptotic bounds (large deviation, moderate deviation, and second-order) can be obtained from those bounds. The marks ✓* indicate that the large deviation bound can be derived up to the critical rate. The computational complexity "Tail" indicates that the computational complexities of those bounds depend on the computational complexities of tail probabilities. It should be noted that Theorem 18 is derived from a special case ($Q_Y = P_Y$) of Theorem 16. The asymptotically optimal choice is $Q_Y = P_Y^{(1+\theta)}$. Under Assumption 1, we can derive the bound of the Markov case only for that special choice of Q_Y, while under Assumption 2, we can derive the bound of the Markov case for the optimal choice of Q_Y. Furthermore, Theorem 18 is not asymptotically tight in the large deviation regime in general, but it is tight if X has no side-information, i.e., the channel is additive. It should be also noted that Theorem 20 does not imply Theorem 18 even for the additive channel case since Assumption 2 restricts the structure of transition matrices even when X has no side-information.

Table 3. Summary of the finite-length bounds for channel coding.

Ach./Conv.	Markov	Single-Shot	$\mathrm{P_c}/\bar{\mathrm{P}}_{\mathrm{c}}$	Complexity	Large Deviation	Moderate Deviation	Second Order
Achievability	Theorem 17 (Ass. 1)	Lemma 22	$\bar{\mathrm{P}}_{\mathrm{c}}$	$O(1)$		✓	
	Theorem 19 (Ass. 2)	Lemma 21	$\bar{\mathrm{P}}_{\mathrm{c}}$	$O(1)$	✓*	✓	
	Theorem 21 (Additive)	Lemma 23	$\bar{\mathrm{P}}_{\mathrm{c}}$	$O(1)$	✓*	✓	
	Lemma 20		$\bar{\mathrm{P}}_{\mathrm{c}}$	Tail		✓	✓
Converse	Theorem 18 (Ass. 1)	(Theorem 16)	$\mathrm{P_c}$	$O(1)$		✓	
	Theorem 20 (Ass. 2)	Theorem 16	$\mathrm{P_c}$	$O(1)$	✓*	✓	
	Theorem 18 (Additive)	(Theorem 16)	$\mathrm{P_c}$	$O(1)$	✓*	✓	
	Lemma 25		$\mathrm{P_c}$	Tail		✓	✓

5. Discussion and Conclusions

In this paper, we developed a unified approach to source coding with side-information and channel coding for a conditional additive channel for finite-length and asymptotic analyses of Markov chains. In our approach, the conditional Rényi entropies defined for transition matrices played important roles. Although we only illustrated the source coding with side-information and the channel coding for a conditional additive channel as applications of our approach, it could be applied to some

other problems in information theory such as random number generation problems, as shown in another paper [60].

Our obtained results for the source coding with side-information and the channel coding of the conditional additive channel has been extended to the case when the side-information is continuous like the real line and the joint distribution X and Y is memoryless. Since this case covers the BPSK-AWGN channel, it can be expected that it covers the MPSK-AWGN channel. Since such channels are often employed in the real channel coding, it is an interesting future topic to investigate the finite-length bound for these channels. Further, we could not define the conditional Rényi entropy for transition matrices of continuous Y. Hence, our result could not be extended to such a continuous case. It is another interesting future topic to extend the obtained result to the case with continuous Y.

Author Contributions: Conceptualization, M.H.; methodology, S.W.; formal analysis, S.W. and M.H.; writing, original draft preparation, S.W.; writing, review and editing, M.H. All authors read and agreed to the published version of the manuscript.

Funding: M.H. is partially supported by the Japan Society of the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (A) No. 23246071, (A) No. 17H01280, (B) No. 16KT0017, the Okawa Research Grant, and Kayamori Foundation of Informational Science Advancement. He is also partially supported by the National Institute of Information and Communication Technology (NICT), Japan. S.W. is supported in part by the Japan Society of the Promotion of Science (JSPS) Grant-in-Aid for Young Scientists (A) No. 16H06091.

Acknowledgments: The authors would like to thank Vincent Y. F. Tan for pointing out Remark 6. The authors are also grateful to Ryo Yaguchi for his helpful comments.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

RCU	random coding union
BSC	binary symmetric channel
DMC	discrete memoryless channel
DT	dependence testing
LDPC	low-density parity check
BPSK	binary phase shift keying
AWGN	additive white Gaussian noise
CGF	cumulant generating function
MPSK	M-ary phase shift keying
CC	channel coding
SC	source coding
SI	side-information

Appendix A. Preparation for the Proofs

When we prove some properties of Rényi entropies or derive converse bounds, some properties of cumulant generating functions (CGFs) become useful. For this purpose, we introduce some terminologies in statistics from [22,23]. Then, in Appendix B, we show the relation between the terminologies in statistics and those in information theory. For the proofs, see [22,23].

Appendix A.1. Single-Shot Setting

Let Z be a random variable with distribution P. Let:

$$\phi(\rho) := \log \mathsf{E}\left[e^{\rho Z}\right] \tag{A1}$$

$$= \log \sum_z P(z) e^{\rho Z} \tag{A2}$$

be the cumulant generating function (CGF). Let us introduce an exponential family:

$$P_\rho(z) := P(z)e^{\rho z - \phi(\rho)}. \tag{A3}$$

By differentiating the CGF, we find that:

$$\begin{aligned} \phi'(\rho) &= \mathsf{E}_\rho[Z] && \text{(A4)} \\ &:= \sum_z P_\rho(z) z. && \text{(A5)} \end{aligned}$$

We also find that:

$$\phi''(\rho) = \sum_z P_\rho(z) \left(z - \mathsf{E}_\rho[Z]\right)^2. \tag{A6}$$

We assume that Z is not constant. Then, (A6) implies that $\phi(\rho)$ is a strict convex function and $\phi'(\rho)$ is monotonically increasing. Thus, we can define the inverse function $\rho(a)$ of $\phi'(\rho)$ by:

$$\phi'(\rho(a)) = a. \tag{A7}$$

Let:

$$D_{1+s}(P\|Q) := \frac{1}{s} \log \sum_z P(z)^{1+s} Q(z)^{-s} \tag{A8}$$

be the Rényi divergence. Then, we have the following relation:

$$sD_{1+s}(P_{\tilde{\rho}}\|P_\rho) = \phi((1+s)\tilde{\rho} - s\rho) - (1+s)\phi(\tilde{\rho}) + s\phi(\rho). \tag{A9}$$

Appendix A.2. Transition Matrix

Let $\{W(z|z')\}_{(z,z')\in\mathcal{Z}^2}$ be an ergodic and irreducible transition matrix, and let $\tilde{P}$ be its stationary distribution. For a function $g : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, let:

$$\mathsf{E}[g] := \sum_{z,z'} \tilde{P}(z')W(z|z')g(z,z'). \tag{A10}$$

We also introduce the following tilted matrix:

$$W_\rho(z|z') := W(z|z')e^{\rho g(z,z')}. \tag{A11}$$

Let λ_ρ be the Perron–Frobenius eigenvalue of W_ρ. Then, the CGF for W with generator g is defined by:

$$\phi(\rho) := \log \lambda_\rho. \tag{A12}$$

Lemma A1. *The function $\phi(\rho)$ is a convex function of ρ, and it is strict convex iff $\phi''(0) > 0$.*

From Lemma A1, $\phi'(\rho)$ is a monotone increasing function. Thus, we can define the inverse function $\rho(a)$ of $\phi'(\rho)$ by:

$$\phi'(\rho(a)) = a. \tag{A13}$$

Appendix A.3. Markov Chain

Let $\mathbf{Z} = \{Z^n\}_{n=1}^{\infty}$ be the Markov chain induced by $W(z|z')$ and an initial distribution P_{Z_1}. For functions $g : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ and $\tilde{g} : \mathcal{Z} \to \mathbb{R}$, let $S_n := \sum_{i=2}^{n} g(Z_i, Z_{i-1}) + \tilde{g}(Z_1)$. Then, the CGF for S_n is given by:

$$\phi_n(\rho) := \log \mathsf{E}\left[e^{\rho S_n}\right]. \tag{A14}$$

We will use the following finite evaluation for $\phi_n(\rho)$.

Lemma A2. *Let v_ρ be the eigenvector of W_ρ^T with respect to the Perron–Frobenius eigenvalue λ_ρ such that $\min_z v_\rho(z) = 1$. Let $w_\rho(z) := P_{Z_1}(z)e^{\rho\tilde{g}(z)}$. Then, we have:*

$$(n-1)\phi(\rho) + \underline{\delta}_\phi(\rho) \le \phi_n(\rho) \le (n-1)\phi(\rho) + \overline{\delta}_\phi(\rho), \tag{A15}$$

where:

$$\overline{\delta}_\phi(\rho) := \log\langle v_\rho | w_\rho\rangle, \tag{A16}$$

$$\underline{\delta}_\phi(\rho) := \log\langle v_\rho | w_\rho\rangle - \log\max_z v_\rho(z). \tag{A17}$$

From this lemma, we have the following.

Corollary A1. *For any initial distribution and $\rho \in \mathbb{R}$, we have:*

$$\lim_{n\to\infty} \phi_n(\rho) = \phi(\rho). \tag{A18}$$

The relation:

$$\lim_{n\to\infty} \frac{1}{n}\mathsf{E}[S_n] = \phi'(0) \tag{A19}$$

$$= \mathsf{E}[g] \tag{A20}$$

is well known. Furthermore, we also have the following.

Lemma A3. *For any initial distribution, we have:*

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Var}\left[S_n\right] = \phi''(0). \tag{A21}$$

Appendix B. Relation Between CGF and Conditional Rényi Entropies

Appendix B.1. Single-Shot Setting

For correlated random variable (X, Y), let us consider $Z = \log \frac{Q_Y(Y)}{P_{XY}(X,Y)}$. Then, the relation between the CGF and conditional Rényi entropy relative to Q_Y is given by:

$$\theta H_{1+\theta}(P_{XY}|Q_Y) = -\phi(-\theta; P_{XY}|Q_Y). \tag{A22}$$

From this, we can also find that the relationship between the inverse functions (cf. (29) and (A7)):

$$\theta(a) = -\rho(a). \tag{A23}$$

Thus, the inverse function defined in (32) also satisfies:

$$(1-\rho(a(R))a(R)+\phi(\rho(a(R));P_{XY}|Q_Y)=R. \tag{A24}$$

Similarly, by setting $Z=\log\frac{1}{P_{X|Y}(X|Y)}$, we have:

$$\theta H^{\downarrow}_{1+\theta}(X|Y)=-\phi(-\theta;P_{XY}|P_Y). \tag{A25}$$

Then, the variance (cf. (11)) satisfies:

$$\mathsf{V}(X|Y)=\phi''(0;P_{XY}|P_Y). \tag{A26}$$

Let $\phi(\rho,\rho')$ be the CGF of $Z=\log\frac{P_Y^{(1-\rho')}(Y)}{P_{XY}(X,Y)}$ (cf. (15) for the definition of $P_Y^{(1-\rho')}$). Then, we have:

$$\theta H_{1+\theta,1+\theta'}(X|Y)=-\phi(-\theta,-\theta'). \tag{A27}$$

It should be noted that $\phi(\rho,\rho')$ is a CGF for fixed ρ', but $\phi(\rho,\rho)$ cannot be treated as a CGF.

Appendix B.2. Transition Matrix

For transition matrix $W(x,y|x',y')$, we consider the function given by:

$$g((x,y),(x',y')):=\log\frac{W(y|y')}{W(x,y|x',y')}. \tag{A28}$$

Then, the relation between the CGF and the lower conditional Rényi entropy is given by:

$$\theta H^{\downarrow,W}_{1+\theta}(X|Y)=-\phi(-\theta). \tag{A29}$$

Then, the variance defined in (51) satisfies:

$$\mathsf{V}^W(X|Y)=\phi''(0). \tag{A30}$$

Appendix C. Proof of Lemma 2

We use the following lemma.

Lemma A4. *For $\theta\in(-1,0)\cup(0,1)$, we have:*

$$H^{\downarrow}_{\frac{1}{1-\theta}}(X|Y)\le H^{\uparrow}_{\frac{1}{1-\theta}}(X|Y)\le H^{\downarrow}_{1+\theta}(X|Y). \tag{A31}$$

Proof. The left hand side inequality of (A31) is obvious from the definition of two Rényi entropies (the latter is defined by taking the maximum). The right-hand side inequality was proven in [71] (Lemma 6). □

Now, we go back to the proof of Lemma 2. From (10) and (11), by the Taylor approximation, we have:

$$H^{\downarrow}_{1+\theta}(X|Y)=H(X|Y)-\frac{1}{2}\mathsf{V}(X|Y)\theta+o(\theta). \tag{A32}$$

Furthermore, since $\frac{1}{1-\theta}=1+\theta+o(\theta)$, we also have:

$$H^{\downarrow}_{\frac{1}{1-\theta}}(X|Y)=H(X|Y)-\frac{1}{2}\mathsf{V}(X|Y)\theta+o(\theta). \tag{A33}$$

Thus, from Lemma A4, we can derive (16) and (17).

Appendix D. Proof of Lemma 3

Statements 1 and 3 follow from the relationships in (A22) and (A25) and the strict convexity of the CGFs.

To prove Statement 5, we first prove the strict convexity of the Gallager function:

$$E_0(\tau; P_{XY}) := \log \sum_y P_Y(y) \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau}} \right)^{1+\tau} \tag{A34}$$

for $\tau > -1$. We use the Hölder inequality:

$$\sum_i a_i^\alpha b_i^\beta \leq \left(\sum_i a_i \right)^\alpha \left(\sum_i b_i \right)^\beta \tag{A35}$$

for $\alpha, \beta > 0$ such that $\alpha + \beta = 1$, where the equality holds iff $a_i = c b_i$ for some constant c. For $\lambda \in (0,1)$, let $1 + \tau_3 = \lambda(1+\tau_1) + (1-\lambda)(1+\tau_2)$, which implies:

$$\frac{1}{1+\tau_3} = \frac{1}{1+\tau_1} \frac{\lambda(1+\tau_1)}{1+\tau_3} + \frac{1}{1+\tau_2} \frac{(1-\lambda)(1+\tau_2)}{1+\tau_3} \tag{A36}$$

and:

$$\frac{\lambda(1+\tau_1)}{1+\tau_3} + \frac{(1-\lambda)(1+\tau_2)}{1+\tau_3} = 1. \tag{A37}$$

Then, by applying the Hölder inequality twice, we have:

$$\sum_y P_Y(y) \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_3}} \right)^{1+\tau_3} \tag{A38}$$

$$= \sum_y P_Y(y) \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_1} \frac{\lambda(1+\tau_1)}{1+\tau_3}} P_{X|Y}(x|y)^{\frac{1}{1+\tau_2} \frac{(1-\lambda)(1+\tau_2)}{1+\tau_3}} \right)^{1+\tau_3} \tag{A39}$$

$$\leq \sum_y P_Y(y) \left[\left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_1}} \right)^{\frac{\lambda(1+\tau_1)}{1+\tau_3}} \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_2}} \right)^{\frac{(1-\lambda)(1+\tau_2)}{1+\tau_3}} \right]^{1+\tau_3} \tag{A40}$$

$$= \sum_y P_Y(y) \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_1}} \right)^{\lambda(1+\tau_1)} \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_2}} \right)^{(1-\lambda)(1+\tau_2)} \tag{A41}$$

$$= \sum_y P_Y(y)^\lambda \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_1}} \right)^{\lambda(1+\tau_1)} P_Y(y)^{1-\lambda} \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_2}} \right)^{(1-\lambda)(1+\tau_2)} \tag{A42}$$

$$\leq \left[\sum_y P_Y(y) \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_1}} \right)^{(1+\tau_1)} \right]^\lambda \left[\sum_y P_Y(y) \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_2}} \right)^{(1+\tau_2)} \right]^{1-\lambda}. \tag{A43}$$

The equality in the second inequality holds iff:

$$\left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_1}} \right)^{1+\tau_1} = c \left(\sum_x P_{X|Y}(x|y)^{\frac{1}{1+\tau_2}} \right)^{1+\tau_2} \quad \forall y \in \mathcal{Y} \tag{A44}$$

for some constant c. Furthermore, the equality in the first inequality holds iff $P_{X|Y}(x|y) = \frac{1}{|\mathrm{supp}(P_{X|Y}(\cdot|y))|}$. Substituting this into (A44), we find that $|\mathrm{supp}(P_{X|Y}(\cdot|y))|$ is irrespective of y. Thus, both the equalities hold simultaneously iff $\mathsf{V}(X|Y) = 0$. Now, since:

$$\theta H^{\uparrow}_{1+\theta}(X|Y) = -(1+\theta) E_0\left(\frac{-\theta}{1+\theta}; P_{XY}\right), \tag{A45}$$

we have:

$$\begin{aligned} \frac{d^2[\theta H^{\uparrow}_{1+\theta}(X|Y)]}{d\theta^2} &= -\frac{1}{(1+\theta)^4} E_0''\left(\frac{-\theta}{1+\theta}; P_{XY}\right) && \text{(A46)} \\ &\leq 0 && \text{(A47)} \end{aligned}$$

for $\theta \in (-1, \infty)$, where the equality holds iff $\mathsf{V}(X|Y) = 0$.

Statement 7 is obvious from the definitions of the two measures. The first part of Statement 8 follows from (A27) and the convexity of the CGF, but we need another argument to check the conditions for strict concavity. Since the second term of:

$$\theta H_{1+\theta,1+\theta'}(X|Y) = -\log \sum_y P_Y(y) \left[\sum_x P_{X|Y}(x|y)^{1+\theta}\right] \left[\sum_x P_{X|Y}(x|y)^{1+\theta'}\right]^{\frac{\theta}{1+\theta'}} + \frac{\theta\theta'}{1+\theta'} H^{\uparrow}_{1+\theta'}(X|Y) \tag{A48}$$

is linear with respect to θ, it suffices to show the strict concavity of the first term. By using the Hölder inequality twice, for $\theta_3 = \lambda\theta_1 + (1-\lambda)\theta_2$, we have:

$$\begin{aligned} &\sum_y P_Y(y) \left[\sum_x P_{X|Y}(x|y)^{1+\theta_3}\right] \left[\sum_x P_{X|Y}(x|y)^{1+\theta'}\right]^{\frac{\theta_3}{1+\theta'}} && \text{(A49)} \\ &\leq \sum_y P_Y(y) \left[\sum_x P_{X|Y}(x|y)^{1+\theta_1}\right]^{\lambda} \left[\sum_x P_{X|Y}(x|y)^{1+\theta_2}\right]^{1-\lambda} \left[\sum_x P_{X|Y}(x|y)^{1+\theta'}\right]^{\frac{\lambda\theta_1+(1-\lambda)\theta_2}{1+\theta'}} && \text{(A50)} \\ &\leq \left[\sum_y P_Y(y) \left[\sum_x P_{X|Y}(x|y)^{1+\theta_1}\right] \left[\sum_x P_{X|Y}(x|y)^{1+\theta'}\right]^{\frac{\theta_1}{1+\theta'}}\right]^{\lambda} && \text{(A51)} \\ &\quad \left[\sum_y P_Y(y) \left[\sum_x P_{X|Y}(x|y)^{1+\theta_2}\right] \left[\sum_x P_{X|Y}(x|y)^{1+\theta'}\right]^{\frac{\theta_2}{1+\theta'}}\right]^{1-\lambda}, && \text{(A52)} \end{aligned}$$

where both the equalities hold simultaneously iff $\mathsf{V}(X|Y) = 0$, which can be proven in a similar manner as the equality conditions in (A40) and (A43). Thus, we have the latter part of Statement 8.

Statements 10–12 are also obvious from the definitions. Statements 2, 4, 6, and 9, follow from Statements 1, 3, 5, and 8, (cf. [71], Lemma 1).

Appendix E. Proof of Lemma 4

Since (24) and (28) are obvious from the definitions, we only prove (26). We note that:

$$\left[\sum_y P_Y(y)\left[\sum_x P_{X|Y}(x|y)^{1+\theta}\right]^{\frac{1}{1+\theta}}\right]^{1+\theta} \tag{A53}$$

$$\leq \left[\sum_y P_Y(y)|\mathrm{supp}(P_{X|Y}(\cdot|y))|^{\frac{1}{1+\theta}}\right]^{1+\theta} \tag{A54}$$

$$\leq \max_{y\in\mathrm{supp}(P_Y)}|\mathrm{supp}(P_{X|Y}(\cdot|y))| \tag{A55}$$

and:

$$\left[\sum_y P_Y(y)\left[\sum_x P_{X|Y}(x|y)^{1+\theta}\right]^{\frac{1}{1+\theta}}\right]^{1+\theta} \tag{A56}$$

$$\geq P_Y(y^*)^{1+\theta}\left[\sum_x P_{X|Y}(x|y^*)^{1+\theta}\right] \tag{A57}$$

$$\overset{\theta\to-1}{\to} |\mathrm{supp}(P_{X|Y}(\cdot|y^*))|, \tag{A58}$$

where:

$$y^* := \operatorname*{argmax}_{y\in\mathrm{supp}(P_Y)}|\mathrm{supp}(P_{X|Y}(\cdot|y))|. \tag{A59}$$

Appendix F. Proof of Lemma 6

From Lemma A4, Theorems 1 and 3, we have:

$$H^{\downarrow,W}_{\frac{1}{1-\theta}}(X|Y) \leq H^{\uparrow,W}_{\frac{1}{1-\theta}}(X|Y) \leq H^{\downarrow,W}_{1+\theta}(X|Y) \tag{A60}$$

for $\theta\in(-1,0)\cup(0,1)$. Thus, we can prove Lemma 6 in the same manner as Lemma 2.

Appendix G. Proof of (63)

First, in the same manner as Theorem 1, we can show:

$$\lim_{n\to\infty}\frac{1}{n}H_{1+\theta}(P_{X^nY^n}|Q_{Y^n}) = H^{W|V}_{1+\theta}(X|Y), \tag{A61}$$

where Q_{Y^n} is a Markov chain induced by V for some initial distribution. Then, since $H_{1+\theta}(P_{X^nY^n}|Q_{Y^n}) \leq H^{\uparrow}_{1+\theta}(X^n|Y^n)$ for each n, by using Theorem 3, we have:

$$H^{W|V}_{1+\theta}(X|Y) \leq H^{\uparrow,W}_{1+\theta}(X|Y). \tag{A62}$$

Thus, the rest of the proof is to show that $H^{\uparrow,W}_{1+\theta}(X|Y)$ is attainable by some V.
Let $\hat{Q}_\theta$ be the normalized left eigenvector of K_θ, and let:

$$V_\theta(y|y') := \frac{\hat{Q}_\theta(y)}{\kappa_\theta\hat{Q}_\theta(y')}K_\theta(y|y'). \tag{A63}$$

Then, V_θ attains the maximum. To prove this, we will show that $\kappa_\theta^{1+\theta}$ is the Perron–Frobenius eigenvalue of:

$$W(x,y|x',y')^{1+\theta}V_\theta(y|y')^{-\theta}. \tag{A64}$$

We first confirm that $(\hat{Q}_\theta(y)^{1+\theta} : (x,y) \in \mathcal{X}\times\mathcal{Y})$ is an eigenvector of (A64) as follows:

$$\sum_{x,y}\hat{Q}_\theta(y)^{1+\theta}W(x,y|x',y')^{1+\theta}V_\theta(y|y')^{-\theta} \tag{A65}$$

$$= \sum_{y}\hat{Q}_\theta(y)^{1+\theta}W_\theta(y|y')\left[\frac{\hat{Q}_\theta(y)}{\kappa_\theta\hat{Q}_\theta(y')}W_\theta(y|y')^{\frac{1}{1+\theta}}\right]^{-\theta} \tag{A66}$$

$$= \kappa_\theta^\theta\hat{Q}_\theta(y')^\theta\sum_{y}\hat{Q}_\theta(y)W_\theta(y|y')^{\frac{1}{1+\theta}} \tag{A67}$$

$$= \kappa_\theta^{1+\theta}\hat{Q}_\theta(y')^{1+\theta}. \tag{A68}$$

Since $(\hat{Q}_\theta(y)^{1+\theta} : (x,y) \in \mathcal{X}\times\mathcal{Y})$ is a positive vector and the Perron–Frobenius eigenvector is the unique positive eigenvector, we find that $\kappa_\theta^{1+\theta}$ is the Perron–Frobenius eigenvalue. Thus, we have:

$$H_{1+\theta}^{W|V_\theta}(X|Y) = -\frac{1+\theta}{\theta}\log\kappa_\theta \tag{A69}$$

$$= H_{1+\theta}^{\uparrow,W}(X|Y). \tag{A70}$$

Appendix H. Proof of Lemma 7

Statement 1 follows from (A29) and the strict convexity of the CGF. Statements 5 and 8–10 follow from the corresponding statements in Lemma 3, Theorems 1, 3 and 4.

Now, we prove (the concavity of $\theta H_{1+\theta}^{\uparrow,W}(X|Y)$ follows from the limiting argument, i.e., the concavity of $\theta H_{1+\theta}^{\uparrow}(X^n|Y^n)$ (cf. Lemma 3) and Theorem 3. However, the strict concavity does not follow from the limiting argument; Statement 3. For this purpose, we introduce the transition matrix counterpart of the Gallager function as follows. Let:

$$\bar{K}_\tau(y|y') := W(y|y')\left[\sum_x W(x|x',y',y)^{\frac{1}{1+\tau}}\right]^{1+\tau} \tag{A71}$$

for $\tau > -1$, which is well defined under Assumption 2. Let $\bar{\kappa}_\tau$ be the Perron–Frobenius eigenvalue of $\bar{K}_\tau$, and let $\tilde{Q}_\tau$ and $\hat{Q}_\tau$ be its normalized right and left eigenvectors. Then, let:

$$L_\tau(y|y') := \frac{\hat{Q}_\tau(y)}{\bar{\kappa}_\tau\hat{Q}_\tau(y')}\bar{K}_\tau(y|y') \tag{A72}$$

be a parametrized transition matrix. The stationary distribution of L_τ is given by:

$$Q_\tau(y') := \frac{\hat{Q}_\tau(y')\tilde{Q}_\tau(y')}{\sum_{y''}\hat{Q}_\tau(y'')\tilde{Q}_\tau(y'')}. \tag{A73}$$

We prove the strict convexity of $E_0^W(\tau) := \log\bar{\kappa}_\tau$ for $\tau > -1$. Then, by the same reason as (A46), we can show Statement 3. Let $Q_\tau(y,y') := L_\tau(y|y')Q_\tau(y')$. By the same calculation as [22] (Proof of Lemmas 13 and 14), we have:

$$\sum_{y,y'}Q_\tau(y,y')\left[\frac{d}{d\tau}\log L_\tau(y|y')\right]^2 = -\sum_{y,y'}Q_\tau(y,y')\left[\frac{d^2}{d\tau^2}\log L_\tau(y|y')\right]. \tag{A74}$$

Furthermore, from the definition of L_τ, we have:

$$-\sum_{y,y'} Q_\tau(y,y') \left[\frac{d^2}{d\tau^2} \log L_\tau(y|y')\right] \tag{A75}$$

$$= -\sum_{y,y'} Q_\tau(y,y') \left[\frac{d^2}{d\tau^2} \log \frac{1}{\kappa_\tau} + \frac{d^2}{d\tau^2} \log \frac{\hat{Q}_\tau(y)}{\hat{Q}_\tau(y')} + \frac{d^2}{d\tau^2} \log K_\tau(y|y')\right] \tag{A76}$$

$$= \frac{d^2}{d\tau^2} \log \kappa_\tau - \sum_{y,y'} Q_\tau(y,y') \frac{d^2}{d\tau^2} \log K_\tau(y|y'). \tag{A77}$$

Now, we show the convexity of $\log \bar{K}_\tau(y|y')$ for each (y,y'). By using the Hölder inequality (cf. Appendix D), for $\tau_3 = \lambda\tau_1 + (1-\lambda)\tau_2$, we have:

$$\left[\sum_x W(x|x',y',y)^{\frac{1}{1+\tau_3}}\right]^{1+\tau_3} \leq \left[\sum_x W(x|x',y',y)^{\frac{1}{1+\tau_1}}\right]^{\lambda(1+\tau_1)} \left[\sum_x W(x|x',y',y)^{\frac{1}{1+\tau_2}}\right]^{(1-\lambda)(1+\tau_2)}. \tag{A78}$$

Thus, $E_0^W(\tau)$ is convex. To check strict convexity, we note that the equality in (A78) holds iff $W(x|x',y',y) = \frac{1}{|\mathrm{supp}(W(\cdot|x',y',y))|}$. Since:

$$\sum_x W(x|x',y',y)^{1+\theta} = \frac{1}{|\mathrm{supp}(W(\cdot|x',y',y))|^\theta} \tag{A79}$$

does not depend on x' from Assumption 2, we have $|\mathrm{supp}(W(\cdot|x',y',y))| = C_{yy'}$ for some integer $C_{yy'}$. By substituting this into $\bar{K}_\tau$, we have:

$$\bar{K}_\tau(y|y') = W(y|y') C_{yy'}^\tau. \tag{A80}$$

On the other hand, we note that the CGF $\phi(\rho)$ is defined as the logarithm of the Perron–Frobenius eigenvalue of:

$$W(x,y|w',y')^{1-\rho} W(y|y')^\rho = W(y|y') \frac{1}{C_{yy'}^{1-\rho}} \mathbf{1}[x \in \mathrm{supp}(W(\cdot|x',y',y))]. \tag{A81}$$

Since:

$$\sum_{x,y} \hat{Q}_\tau(y) W(y|y') \frac{1}{C_{yy'}^{1-\tau}} \mathbf{1}[x \in \mathrm{supp}(W(\cdot|x',y',y))] \tag{A82}$$

$$= \sum_y \hat{Q}_\tau(y) W(y|y') C_{yy'}^\tau \tag{A83}$$

$$= \bar{\kappa}_\tau \hat{Q}_\tau(y'), \tag{A84}$$

$\bar{\kappa}_\tau$ is the Perron–Frobenius eigenvalue of (A81), and thus, we have $E_0^W(\tau) = \phi(\tau)$ when the equality in (A78) holds for every (y,y') such that $W(y|y') > 0$. Since $\phi(\tau)$ is strict convex if $\mathrm{V}^W(X|Y) > 0$, $E_0^W(\tau)$ is strict convex if $\mathrm{V}^W(X|Y) > 0$. Thus, $\theta H_{1+\theta}^{\uparrow,W}(X|Y)$ is strict concave if $\mathrm{V}^W(X|Y) > 0$. On the other hand, from (57), $\theta H_{1+\theta}^{\uparrow,W}(X|Y)$ is strict concave only if $\mathrm{V}^W(X|Y) > 0$.

Statement 6 can be proven by modifying the proof of Statement 8 of Lemma 3 to a transition matrix in a similar manner as Statement 3 of the present lemma.

Finally, Statements 2, 4 and 7 follow from Statements 1, 3 and 6 (cf. [71], Lemma 1).

Appendix I. Proof of Lemma 9

We only prove (75) since we can prove (76) exactly in the same manner by replacing $H_{1+\theta}^{\downarrow,W}(X|Y)$, $\theta^{\downarrow}(a)$, and $a^{\downarrow}(R)$ by $H_{1+\theta}^{\uparrow,W}(X|Y)$, $\theta^{\uparrow}(a)$, and $a^{\uparrow}(R)$. Let:

$$f(\theta) := \frac{-\theta R + \theta H_{1+\theta}^{\downarrow,W}(X|Y)}{1+\theta}. \tag{A85}$$

Then, we have:

$$f'(\theta) = \frac{-R + (1+\theta)\frac{d[\theta H_{1+\theta}^{\downarrow,W}(X|Y)]}{d\theta} - \theta H_{1+\theta}^{\downarrow,W}(X|Y)}{(1+\theta)^2} \tag{A86}$$

$$= \frac{-R + R\left(\frac{d[\theta H_{1+\theta}^{\downarrow,W}(X|Y)]}{d\theta}\right)}{(1+\theta)^2}. \tag{A87}$$

Since $R(a)$ is monotonically increasing and $\frac{d[\theta H_{1+\theta}^{\downarrow,W}(X|Y)]}{d\theta}$ is monotonically decreasing, we have $f'(\theta) \geq 0$ for $\theta \leq \theta(a(R))$ and $f'(\theta) \leq 0$ for $\theta \geq \theta(a(R))$. Thus, $f(\theta)$ takes its maximum at $\theta(a(R))$. Furthermore, since $-1 \leq \theta(a(R)) \leq 0$ for $H^W(X|Y) \leq R \leq H_0^{\downarrow,W}(X|Y)$, we have:

$$\sup_{-1\leq\theta\leq 0} \frac{-\theta R + \theta H_{1+\theta}^{\downarrow,W}(X|Y)}{1+\theta} \tag{A88}$$

$$= \frac{-\theta(a(R))R + \theta(a(R))H_{1+\theta(a(R))}^{\downarrow,W}(X|Y)}{1+\theta(a(R))} \tag{A89}$$

$$= \frac{-\theta(a(R))[(1+\theta(a(R)))a(R) - \theta(a(R))H_{1+\theta(a(R))}^{\downarrow,W}(X|Y)] + \theta(a(R))H_{1+\theta(a(R))}^{\downarrow,W}(X|Y)}{1+\theta(a(R))} \tag{A90}$$

$$= -\theta(a(R))a(R) + \theta(a(R))H_{1+\theta(a(R))}^{\downarrow,W}(X|Y), \tag{A91}$$

where we substituted $R = R(a(R))$ in the second equality.

Appendix J. Proof of Lemma 11

Let u be the vector such that $u(y) = 1$ for every $y \in \mathcal{Y}$. From the definition of $H_{1+\theta}^{\uparrow}(X^n|Y^n)$, we have the following sequence of calculations:

$$e^{-\frac{\theta}{1+\theta}\theta H_{1+\theta}^{\uparrow}(X^n|Y^n)} \tag{A92}$$

$$= \sum_{y_1,\ldots,y_n}\left[\sum_{x_n,\ldots,x_1} P(x_1,y_1)^{1+\theta}\prod_{i=2}^{n} W(x_i,y_i|x_{i-1},y_{i-1})^{1+\theta}\right]^{\frac{1}{1+\theta}} \tag{A93}$$

$$\overset{(a)}{=} \sum_{y_n,\ldots,y_1}\left[\sum_{x_1} P(x_1,y_1)^{1+\theta}\right]^{\frac{1}{1+\theta}}\prod_{i=2}^{n} W_\theta(y_i|y_{i-1})^{\frac{1}{1+\theta}} \tag{A94}$$

$$= \langle u|K_\theta^{n-1}w_\theta\rangle \tag{A95}$$

$$\leq \langle v_\tau|K_\theta^{n-1}w_\theta\rangle \tag{A96}$$

$$= \langle (K_\theta^T)^{n-1}v_\theta|w_\theta\rangle \tag{A97}$$

$$= \kappa_\theta^{n-1}\langle v_\theta|w_\theta\rangle \tag{A98}$$

$$= e^{-(n-1)\frac{\theta}{1+\theta}H_{1+\theta}^{\uparrow,W}(X|Y)}\langle v_\theta|w_\theta\rangle, \tag{A99}$$

which implies the left-hand side inequality, where we used Assumption 2 in (a). On the other hand, we have the following sequence of calculations:

$$
\begin{aligned}
& e^{-\frac{\theta}{1+\theta}\theta H^{\uparrow}_{1+\theta}(X^n|Y^n)} && \text{(A100)}\\
&= \langle u|K_\theta^{n-1}w_\theta\rangle && \text{(A101)}\\
&\geq \frac{1}{\max_y v_\theta(y)}\langle v_\theta|K_\theta^{n-1}w_\theta\rangle && \text{(A102)}\\
&= \frac{1}{\max_y v_\theta(y)}\langle (K_\theta^T)^{n-1}v_\theta|w_\theta\rangle && \text{(A103)}\\
&= \kappa_\theta^{n-1}\frac{\langle v_\theta|w_\theta\rangle}{\max_y v_\theta(y)} && \text{(A104)}\\
&= e^{-(n-1)\frac{\theta}{1+\theta}H^{\uparrow,W}_{1+\theta}(X|Y)}\frac{\langle v_\theta|w_\theta\rangle}{\max_y v_\theta(y)}, && \text{(A105)}
\end{aligned}
$$

which implies the right-hand side inequality.

Appendix K. Proof of Theorem 5

For arbitrary $\tilde{\rho}\in\mathbb{R}$, we set $\alpha := P_{XY}\{X\neq \mathsf{d}(\mathsf{e}(X),Y)\}$ and $\beta := P_{XY,\tilde{\rho}}\{X\neq \mathsf{d}(\mathsf{e}(X),Y)\}$, where:

$$
P_{XY,\rho}(x,y) := P_{XY}(x,y)^{1-\rho}Q_Y(y)^\rho e^{-\phi(\rho;P_{XY}|Q_Y)}. \tag{A106}
$$

Then, by the monotonicity of the Rényi divergence, we have:

$$
\begin{aligned}
sD_{1+s}(P_{XY,\tilde{\rho}}\|P_{XY}) &\geq \log\left[\beta^{1+s}\alpha^{-s}+(1-\beta)^{1+s}(1-\alpha)^{-s}\right] && \text{(A107)}\\
&\geq \log\beta^{1+s}\alpha^{-s}. && \text{(A108)}
\end{aligned}
$$

Thus, we have:

$$
-\log\alpha \leq \frac{\phi((1+s)\tilde{\rho};P_{XY}|Q_Y)-(1+s)\phi(\tilde{\rho};P_{XY}|Q_Y)-(1+s)\log\beta}{s}. \tag{A109}
$$

Now, by using Lemma 18, we have:

$$
1-\beta \leq P_{XY,\tilde{\rho}}\left\{\log\frac{Q_Y(y)}{P_{XY,\tilde{\rho}}(x,y)}\leq\gamma\right\}+\frac{M}{e^\gamma}. \tag{A110}
$$

We also have, for any $\sigma\leq 0$,

$$
\begin{aligned}
& P_{XY,\tilde{\rho}}\left\{\log\frac{Q_Y(y)}{P_{XY,\tilde{\rho}}(x,y)}\leq\gamma\right\} && \text{(A111)}\\
&\leq \sum_{x,y}P_{XY,\tilde{\rho}}(x,y)e^{\sigma\left(\log\frac{Q_Y(y)}{P_{XY,\tilde{\rho}}(x,y)}-\gamma\right)} && \text{(A112)}\\
&= e^{-[\sigma\gamma-\phi(\sigma;P_{XY,\tilde{\rho}}|Q_Y)]}. && \text{(A113)}
\end{aligned}
$$

Thus, by setting γ so that:

$$
\sigma\gamma-\phi(\sigma;P_{XY,\tilde{\rho}}|Q_Y)=\gamma-R, \tag{A114}
$$

we have

$$1-\beta \le 2e^{-\frac{\sigma R-\phi(\sigma;P_{XY,\tilde{\rho}}|Q_Y)}{1-\sigma}}. \tag{A115}$$

Furthermore, we have the relation:

$$\begin{aligned}
\phi(\sigma;P_{XY,\tilde{\rho}}|Q_Y) &= \log\sum_{x,y}P_{XY,\tilde{\rho}}(x,y)^{1-\sigma}Q_Y(y)^{\sigma} && \text{(A116)}\\
&= \log\sum_{x,y}\left(P_{XY}(x,y)^{1-\tilde{\rho}}Q_Y(y)^{\tilde{\rho}}e^{-\phi(\tilde{\rho};P_{XY}|Q_Y)}\right)^{1-\sigma}Q_Y(y)^{\sigma} && \text{(A117)}\\
&= -(1-\sigma)\phi(\tilde{\rho};P_{XY}|Q_Y)+\log\sum_{x,y}P_{XY}(x,y)^{1-\tilde{\rho}-\sigma(1-\tilde{\rho})}Q_Y(y)^{\tilde{\rho}+\sigma(1-\tilde{\rho})} && \text{(A118)}\\
&= \phi(\tilde{\rho}+\sigma(1-\tilde{\rho});P_{XY}|Q_Y)-(1-\sigma)\phi(\tilde{\rho};P_{XY}|Q_Y). && \text{(A119)}
\end{aligned}$$

Thus, by substituting $\tilde{\rho}=-\tilde{\theta}$ and $\sigma=-\vartheta$ and by using (A22), we can derive (124).
Now, we restrict the range of $\tilde{\rho}$ so that $\rho(a(R))<\tilde{\rho}<1$ and take:

$$\sigma=\frac{\rho(a(R))-\tilde{\rho}}{1-\tilde{\rho}}. \tag{A120}$$

Then, by substituting this into (A119) and (A119) into (A115), we have ($\phi(\rho;P_{XY}|Q_Y)$ is omitted as $\phi(\rho)$):

$$\begin{aligned}
&\frac{\sigma R-\phi(\tilde{\rho}+\sigma(1-\tilde{\rho}))+(1-\sigma)\phi(\tilde{\rho})}{1-\sigma} && \text{(A121)}\\
&= \frac{(\rho(a(R))-\tilde{\rho})R-(1-\tilde{\rho})\phi(\rho(a(R)))+(1-\rho(a(R)))\phi(\tilde{\rho})}{1-\rho(a(R))} && \text{(A122)}\\
&= \frac{(\rho(a(R))-\tilde{\rho})\{(1-\rho(a(R)))a(R)+\phi(\rho(a(R)))\}-(1-\tilde{\rho})\phi(\rho(a(R)))+(1-\rho(a(R)))\phi(\tilde{\rho})}{1-\rho(a(R))} && \text{(A123)}\\
&= (\rho(a(R))-\tilde{\rho})a(R)-\phi(\rho(a(R)))+\phi(\tilde{\rho}), && \text{(A124)}
\end{aligned}$$

where we used (A24) in the second equality. Thus, by substituting $\tilde{\rho}=-\tilde{\theta}$ and by using (A22) again, we have (125).

Appendix L. Proof of Theorem 16

Let:

$$P_{X^nY^n,\rho}(x^n,y^n):=P_{X^nY^n}(x^n,y^n)^{1-\rho}Q_{Y^n}(y^n)^{\rho}e^{-\phi(\rho;P_{X^nY^n}|Q_{Y^n})}, \tag{A125}$$

and let $P_{B^n|A^n,\rho}$ be a conditional additive channel defined by:

$$P_{B^n|A^n,\rho}(a^n+x^n|a^n)=P_{X^nY^n,\rho}(x^n,y^n). \tag{A126}$$

We also define the joint distribution of the message, the input, the output, and the decoded message for each channel:

$$\begin{aligned}
P_{M_nA^nB^n\hat{M}_n}(m,a^n,b^n,\hat{m}) &:= \frac{1}{M_n}\mathbf{1}[\mathsf{e}_n(m)=a^n]P_{B^n|A^n}(b^n|a^n)\mathbf{1}[\mathsf{d}_n(b^n)=\hat{m}], && \text{(A127)}\\
P_{M_nA^nB^n\hat{M}_n,\rho}(m,a^n,b^n,\hat{m}) &:= \frac{1}{M_n}\mathbf{1}[\mathsf{e}_n(m)=a^n]P_{B^n|A^n,\rho}(b^n|a^n)\mathbf{1}[\mathsf{d}_n(b^n)=\hat{m}]. && \text{(A128)}
\end{aligned}$$

For arbitrary $\tilde{\rho} \in \mathbb{R}$, let $\alpha := P_{M_n\hat{M}_n}\{m \neq \hat{m}\}$ and $\beta := P_{M_n\hat{M}_n,\tilde{\rho}}\{m \neq \hat{m}\}$. Then, by the monotonicity of the Rényi divergence, we have:

$$\begin{aligned} sD_{1+s}(P_{A^nB^n,\tilde{\rho}}\|P_{A^nB^n}) &\geq sD_{1+s}(P_{M_n\hat{M}_n,\tilde{\rho}}\|P_{M_n\hat{M}_n}) && \text{(A129)} \\ &\geq \log\left[\beta^{1+s}\alpha^{-s} + (1-\beta)^{1+s}(1-\alpha)^{-s}\right] && \text{(A130)} \\ &\geq \log\beta^{1+s}\alpha^{-s}. && \text{(A131)} \end{aligned}$$

Thus, we have:

$$-\log\alpha \leq \frac{sD_{1+s}(P_{A^nB^n,\tilde{\rho}}\|P_{A^nB^n}) - (1+s)\log\beta}{s}. \tag{A132}$$

Here, we have:

$$D_{1+s}(P_{A^nB^n,\tilde{\rho}}\|P_{A^nB^n}) = D_{1+s}(P_{X^nY^n,\tilde{\rho}}\|P_{X^nY^n}). \tag{A133}$$

On the other hand, from Lemma 25, we have:

$$1-\beta \leq P_{X^nY^n,\tilde{\rho}}\left\{\log\frac{Q_{Y^n}(y^n)}{P_{X^nY^n,\tilde{\rho}}(x^n,y^n)} \leq n\log|\mathcal{A}| - \gamma\right\} + \frac{e^R}{e^{n\log|\mathcal{A}|-\gamma}}. \tag{A134}$$

Thus, by the same argument as in (A111)–(A119) and by noting (A22), we can derive (222). Now, we restrict the range of $\tilde{\rho}$ so that $\rho(a(R)) < \tilde{\rho} < 1$ and take:

$$\sigma = \frac{\rho(a(R)) - \tilde{\rho}}{1-\tilde{\rho}}. \tag{A135}$$

Then, by noting (A22), we have (223).

References

1. Polyanskiy, Y.; Poor, H.V.; Verdu, S. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory* **2010**, *56*, 2307–2359. [CrossRef]
2. Verdú, S.; Han, T.S. A general fomula for channel capacity. *IEEE Trans. Inform. Theory* **1994**, *40*, 1147–1157. [CrossRef]
3. Han, T.S. *Information-Spectrum Methods in Information Theory*; Springer: Berlin/Heidelberg, Germany, 2003.
4. Hayashi, M.; Nagaoka, H. General formulas for capacity of classical-quantum channels. *IEEE Trans. Inf. Theory* **2003**, *49*, 1753–1768. [CrossRef]
5. Wang, L.; Renner, R. One-shot classical-quantum capacity and hypothesis testing. *Phys. Rev. Lett.*, **2012**, *108*, 200501. [CrossRef] [PubMed]
6. Gallager, R.G. A simple derivation of the coding theorem and some applications. *IEEE Trans. Inf. Theory* **1965**, *11*, 3–18. [CrossRef]
7. Polyanskiy, Y. Channel coding: Non-Asymptotic Fundamental Limits. Ph.D. Dissertation, Princeton University, Princeton, NJ, USA, November 2010.
8. Tomamichel, M.; Hayashi, M. A hierarchy of information quantities for finite block length analysis of quantum tasks. *IEEE Trans. Inform. Theory* **2013**, *59*, 7693–7710. [CrossRef]
9. Matthews, W.; Wehner, S. Finite blocklength converse bounds for quantum channels. *IEEE Trans. Inf. Theory* **2014**, *60*, 7317–7329. [CrossRef]
10. Gallager, R.G. *Information Theory and Reliable Communication*; John Wiley & Sons: Hoboken, NJ, USA, 1968.
11. Hayashi, M. Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. Inf. Theory* **2009**, *55*, 4947–4966. [CrossRef]
12. Hayashi, M. Second-order asymptotics in fixed-length source coding and intrinsic randomness. *IEEE Trans. Inf. Theory* **2008**, *54*, 4619–4637. [CrossRef]

13. Altug, Y.; Wagner, A.B. Moderate deviation analysis of channel coding: Discrete memoryless case. In Proceedings of the IEEE International Symposium on Information Theory, Austin, TX, USA, 13–18 June 2010; pp. 265–269.
14. He, D.; Lastras-Montano, L.A.; Yang, E.; Jagmohan, A.; Chen, J. On the redundancy of slepian-wolf coding. *IEEE Trans. Inf. Theory* **2009**, *55*, 5607–5627. [CrossRef]
15. Tan, V.Y.F. Moderate-deviations of lossy source coding for discrete and gaussian sources. In Proceedings of the 2012 IEEE International Symposium on Information Theory, Cambridge, MA, USA, 1–6 July 2012; pp. 920–924.
16. Kuzuoka, S. A simple technique for bounding the redundancy of source coding with side-information. In Proceedings of the 2012 IEEE International Symposium on Information Theory, Cambridge, MA, USA, 1–6 July 2012; pp. 915–919.
17. Yang, E.; Meng, J. New nonasymptotic channel coding theorems for structured codes. *IEEE Trans. Inf. Theory* **2015**, *61*, 4534–4553. [CrossRef]
18. Arimoto, S. Information measures and capacity of order α for discrete memoryless channels. In *Colloquia Mathematica Societatis Janos Bolyai, 16. Topics in Information Theory*; Elsevier: Amsterdam, The Netherlands, 1975; pp. 41–52.
19. Hayashi, M. Exponential decreasing rate of leaked information in universal random privacy amplification. *IEEE Trans. Inf. Theory* **2011**, *57*, 3989–4001. [CrossRef]
20. Teixeira, A.; Matos, A.; Antunes, L. Conditional Rényi entropies. *IEEE Trans. Inf. Theory* **2012**, *58*, 4273–4277. [CrossRef]
21. Iwamoto, M.; Shikata, J. Information theoretic security for encryption based on conditional Rényi entropies. In *Information Theoretic Security ICITS 2013*; Padró, C., Ed.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2013; pp. 103–121.
22. Hayashi, M.; Watanabe, S. Information geometry approach to parameter estimation in Markov chains. *Ann. Stat.* **2016**, *44*, 1495–1535. [CrossRef]
23. Watanabe, S.; Hayashi, M. Finite-length analysis on tail probability and simple hypothesis testing for Markov chain. *Ann. Appl. Probab.* **2017**, *27*, 811–845. [CrossRef]
24. Wyner, A.D. Recent results in the shannon theory. *IEEE Trans. Inf. Theory* **1974**, *20*, 2–10. [CrossRef]
25. Csiszár, I. Linear codes for sources and source networks: Error exponents, universal coding. *IEEE Trans. Inf. Theory* **1982**, *28*, 585–592. [CrossRef]
26. Ahlswede, R.; Dueck, G. Good codes can be produced by a few permutations. *IEEE Trans. Inf. Theory* **1982**, *28*, 430–443. [CrossRef]
27. Chen, J.; He, D.-K.; Jagmohan, A.; Lastras-Montano, L.A.; Yang, E. On the linear codebook-level duality between Slepian-Wolf coding and channel coding. *IEEE Trans. Inf. Theory* **2009**, *55*, 5575–5590. [CrossRef]
28. Hayashi, M. Tight exponential analysis of universally composable privacy amplification and its applications. *IEEE Trans. Inf. Theory* **2013**, *59*, 7728–7746. [CrossRef]
29. Gilbert, E.N. Capacity of burst-noise for codes on burst-noise channels. *Bell Syst. Tech. J.* **1960**, *39*, 1253–1265. [CrossRef]
30. Elliott, E.O. Estimates of error rates for codes on burst-noise channels. *Bell Syst. Tech. J.* **1963**, *42*, 1977–1997. [CrossRef]
31. Delsarte, P.; Piret, P. Algebraic construction of shannon codes for regular channels. *IEEE Trans. Inf. Theory* **1982**, *28*, 593–599. [CrossRef]
32. Tomamichel, M.; Tan, V.Y.F. Second-order coding rates for channels with state. *IEEE Trans. Inf. Theory* **2014**, *60*, 4427–4448. [CrossRef]
33. Kemeny, J.G.; Snell, J. *Finite Markov Chains*; Springer: Berlin/Heidelberg, Germany, 1976.
34. Feller, W. *An Introduction to Probability Theory and Its Applications*; Wiley: Hoboken, NJ, USA, 1971.
35. Tikhomirov, A.N. On the convergence rate in the central limit theorem for weakly dependent random variables. *Theory Probab. Appl.* **1980**, *25*, 790–890. [CrossRef]
36. Kontoyiannis, I.; Meyn, P. Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.* **2003**, *13*, 304–362. [CrossRef]
37. Hervé, L.; Ledoux, J.; Patilea, V. A uniform Berry-Esseen theorem on *m*-estimators for geometrically ergodic Markov chains. *Bernoulli* **2012**, *18*, 703–734. [CrossRef]

38. Watanabe, S.; Hayashi, M. Non-asymptotic analysis of privacy amplification via Rényi entropy and inf-spectral entropy. In Proceedings of the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, 7–12 July 2013; pp. 2715–2719.
39. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006.
40. Davisson, L.D.; Longo, G.; Sgarro, A. The error exponent for the noiseless encoding of finite ergodic Markov sources. *IEEE Trans. Inform. Theory* **1981**, *27*, 431–438. [CrossRef]
41. Vašek, K. On the error exponent for ergodic Markov source. *Kybernetika* **1980**, *16*, 318–329.
42. Zhong, Y.; Alajaji, F.; Campbell, L.L. Joint source-channel coding error exponent for discrete communication systems with Markovian memory. *IEEE Trans. Inf. Theory* **2007**, *53*, 4457–4472. [CrossRef]
43. Polyanskiy, Y.; Poor, H.V.; Verdu, S. Dispersion of the Gilbert-Elliott channel. *IEEE Trans. Inf. Theory* **2011**, *57*, 1829–1848. [CrossRef]
44. Kontoyiannis, I. Second-order noiseless source coding theorems. *IEEE Trans. Inform. Theory* **1997**, *43*, 1339–1341. [CrossRef]
45. Kontoyiannis, I.; Verdú, S. Optimal lossless data compression: Non-asymptotic and asymptotics. *IEEE Trans. Inf. Theory* **2014**, *60*, 777–795. [CrossRef]
46. Scarlett, J.; Martinez, A.; Fábregas, A.G.i. Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations. *IEEE Trans. Inform. Theory* **2014**, *60*, 2647–2666. [CrossRef]
47. Scarlett, J.; Martinez, A.; i Fabregas, A.G. The saddlepoint approximation: A unification of exponents, dispersions and moderate deviations. *arXiv* **2014**, arXiv:1402.3941.
48. Ben-Ari, I.; Neumann, M.. Probabilistic approach to Perron root, the group inverse, and applications. *Linear Multilinear Algebra* **2010**, *60*, 39–63. [CrossRef]
49. Lalley, S.P. Ruelle's Perron-Frobenius theorem and the central limit theorem for additive functionals of one-dimensional Gibbs states. *Adapt. Stat. Proced. Relat. Top.* **1986**, *8*, 428–446.
50. Kato, T. *Perturbation Theory for Linear Operators*; Springer: New York, NY, USA, 1980.
51. Häggström, O.; Rosenthal, J.S. On the central limit theorem for geometrically ergodic markov chains. *Electron. Commun. Probab.* **2007**, *12*, 454–464.
52. Kipnis, C.; Varadhan, S.R.S. Central limit theorem for additive functionals of reversible markov processes and applications to simple exclusions. *Commun. Math. Phys.* **1986**, *104*, 1–19. [CrossRef]
53. Komorowski, C.L.T.; Olla, S. *Fluctuations in Markov Processes: Time Symmetry and Martingale Approximation*; Springer: Berlin, Germany, 2012.
54. Meyn, S.P.; Tweedie, R.L. *Markov Chains and Stochastic Stability*; Springer: London, UK, 1993.
55. Jones, G.L. On the Markov chain central limit theorem. *Probab. Surv.* **2004**, *1*, 299–320, . [CrossRef]
56. Tomamichel, M.; Berta, M.; Hayashi, M. Relating different quantum generalizations of the conditional Rényi entropy. *J. Math. Phys.* **2014**, *55*, 082206. [CrossRef]
57. Hayashi, M. Large deviation analysis for classical and quantum security via approximate smoothing. *IEEE Trans. Inf. Theory* **2014**, *60*, 6702–6732. [CrossRef]
58. Csiszár, I. Generalized cutoff rates and Rényi's information measures. *IEEE Trans. Inf. Theory* **1995**, *41*, 6–34. [CrossRef]
59. Polyanskiy, Y.; Verdú, S. Arimoto channel coding converse and Rényi divergence. In Proceedings of the 2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Allerton, IL, USA, 29 September–1 October 2010; pp. 1327–1333.
60. Hayashi, M.; Watanabe, S. Uniform random number generation from Markov chains: Non-asymptotic and asymptotic analyses. *IEEE Trans. Inform. Theory* **2016**, *62*, 1795–1822. [CrossRef]
61. Kemeny, J.G.; Snell, J.L. *Finite Markov Chains*; Springer: New York, NY, USA, 1960.
62. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Express: Cambridge, UK, 1985.
63. Wegman, M.N.; Carter, J.L. New hash functions and their use in authentication and set equality. *J. Comput. Syst. Sci.* **1981**, *22*, 265–279. [CrossRef]
64. Gallager, R.G. Source coding with side-information and universal coding. *Proc. IEEE Int. Symp. Inf. Theory* **1976**. Available online: http://web.mit.edu/gallager/www/papers/paper5.pdf (accessed on 5 April 2020).
65. Hayashi, M. *Quantum Information: An Introduction*; Springer: Berlin/Heidelberg, Germany, 2006.
66. Renner, R.; Wolf, S. Simple and tight bound for information reconciliation and privacy amplification. In *Advances in Cryptology – ASIACRYPT 2005*, ser. Lecture Notes in Computer Science, Springer: Berlin/Heidelberg, Germany, 2005; pp. 199–216.

67. Billingsley, P.; *Probability and Measure*; JOHN WILEY & SONS: Hoboken, NJ, USA, 1995.
68. Cover, T. A proof of the data compression theorem of Slepian and Wolf for ergodic sources. *IEEE Trans. Inf. Theory* **1975**, *21*, 226–228. [CrossRef]
69. Dembo, A.; Zeitouni, O. *Large Deviations Techniques and Applications*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1998.
70. Hayashi, M. Error exponent in asymmetric quantum hypothesis testing and its application to classical-quantum channel coding. *Phys. Rev. A* **2007**, *76*, 062301. [CrossRef]
71. Hayashi, M. Security analysis of ε-almost dual universal$_2$ hash functions. *IEEE Trans. Inf. Theory* **2016**, *62*, 3451–3476. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).